This is a reproduction of a library book that was digitized by Google as part of an ongoing effort to preserve the information in books and make it universally accessible.

Google™ books

https://books.google.com



Digitized by Google

Digitized by Google

Digitized by Google

Digitized by Google

Digitized by Google

Digitized by Google

Digitized by Google

Digitized by Google

INTRODUCTION TO THE MATHEMATICS OF STATISTICS

BY

ROBERT WILBUR BURGESS, PH.D.

Senior Statistician, Western Electric Company
Formerly Assistant Professor of Mathematics, Brown University

UNDER THE EDITORSHIP OF

JOHN WESLEY YOUNG

Cheney Professor of Mathematics, Dartmouth College



HOUGHTON MIFFLIN COMPANY
BOSTON · NEW YORK · CHICAGO · DALLAS · SAN FRANCISCO
The Riverside Press Cambridge

Digitized by Google

COPYRIGHT, 1927, BY ROBERT WILBUR BURGESS

ALL RIGHTS RESERVED

replacing M512697

GIFT

The Riverside Press

CAMBRIDGE • MASSACHUSETTS

PRINTED IN THE U.S.A.

Digitized by Google

HA29
B8

EDITOR'S INTRODUCTION

THE more serious books on statistics which have appeared of late in this country fall essentially into two classes: those written primarily by mathematicians, which are largely theoretical and make considerable demands on the mathematical preparation of their readers; and those written by specialists in other fields intended primarily for use by other specialists in these fields. The latter are usually limited as to their methods and are often unsound as to their mathematics.

The present text differs from both of these classes. It presupposes as to mathematical preparation only the usual secondary-school mathematics. It is written by a man of wide experience both as a teacher of mathematics and as a practical statistician, and it is general in its scope. The distinguishing features of the text are set forth in more detail in the author's preface. Suffice it to say here that the book should fill a real need on the part of those who desire a thorough general training in the mathematics of statistics. It would seem to be well adapted not merely for use as a textbook in college courses, but also for self-study on the part of those who desire to obtain a thoroughgoing working knowledge of statistical methods.

J. W. YOUNG

M703043

Digitized by Google

Digitized by Google

PREFACE

General aim of this book. A rather vaguely defined thing called "statistics" is being demanded with increasing emphasis in various branches of human activity. The name "statistics" reflects an origin connected with the State; at the present time, however, the banker, railroad official, educator, economist, or public-health worker is as likely as the public official to insist on the need for more accurate quantitative data, and more searching and refined methods in analyzing these data.

The aim of the following pages is to set forth a general but elementary outline of the best methods of statistical analysis, with an explanation and discussion of these methods as complete and as thorough as is permitted by the space available and by the previous training which may be expected of the reader.

Distinguishing features of the treatment. The treatment here given will differ in one respect or another from almost every other discussion of this general subject. Important distinguishing characteristics are as follows:

(1) The treatment is general rather than related to one particular field, such as education, biometry, economics, or business administration. The growth of peculiar methods in any one field is, we feel, wasteful and confusing. Some fields, to be sure, require more emphasis on certain developments of the subject than is normally required in others; but the statistician in any field will find occasion to use almost every known method, if he really assails the problems which confront him. It is clear, therefore, that we should not restrict the availability of a statistical method by expressing it in an unusual terminology, or by departing from accepted statistical practice in minor details.

(2) The treatment is mathematical in the sense that it is the orderly and logical discussion of the methods used with quantitative data of certain types, and that it is not concerned, except incidentally, with the methods of gathering the individual quantita-

Digitized by Google

tive facts, or with the mechanical details of tabulation or graphing, or with the various branches of science or business which use statistical methods.

(3) The treatment, even in its mathematical aspects, is very much simplified. Many men and women who really need a knowledge of statistics are convinced that they find mathematics hard. To assist such persons, and indeed all readers, to get that thorough grasp of the points we shall discuss which is essential to their being of real value, we shall parallel all algebraic proofs with numerical illustrations and, where possible, with graphic and verbal discussion. Each of these methods of presentation has its advantages: the numerical illustrations, although only special cases, will appeal to most readers, even the mathematical, as making the point clearer; the algebraic method yields a real proof and facilitates general statements as to limitations, range of applicability of the principle, etc.; the graphic method produces a summary easy to carry in the mind's eye; and the verbal statement is usually easiest to apply as a working rule. It is recommended, therefore, that the reader try to understand all four methods of approach. If he follows this course in the earlier chapters, he will find that the later chapters will not seem so difficult as he may have expected.

(4) A careful discussion is given of mathematical principles that are often ignored as not worthy of attention. These are the essential principles back of much statistical work in percentages, rates, and averages, which are usually thought to involve no serious mathematics. These principles are observed more or less unconsciously by good statisticians; the object in treating them so fully here is to facilitate their application by making them explicit, and to enable beginners to grasp them firmly at the start, rather than merely to grow into them. It will be found, moreover, that an explicit understanding of these principles provides a sound basis for a discussion of questions raised by the more advanced methods.

(5) The more complicated mathematical concepts — measures of dispersion, regression equations, probability, frequency curves, and correlation — are introduced by discussions from the common-sense point of view.

(6) Graphic methods are not treated in a separate chapter, but

Digitized by Google

each method is taken up along with the discussion of the principle which it is used to illustrate. The point of view of this book is that some real thinking must underlie any graphic presentation to make it significant, and that logically the emphasis belongs on that thinking rather than on the graphic device.

For whom is the book intended? It is hoped that the book may be useful to the following classes of actual or prospective statistical workers:

(1) College students of statistics, mathematics, economics, education, and other subjects in which statistical methods are helpful.

(2) Practicing business statisticians who desire to learn something of the methods of regression and correlation, as well as to systematize their theoretical understanding of more elementary fields.

(3) Accountants who have occasional need to enter the field of statistics.

(4) Teachers of mathematics who desire to be able to meet more fully the growing demand for statistical work of an elementary sort in high-school or first-year college classes.

(5) Specialists in education, economics, public-health work, and other fields, who can use to advantage a statement of mathematical principles in connection with the statistics of their own special work.

In order to meet as far as possible the needs of all these groups, all mathematics beyond that given in a good high-school course is explained before it is used, or reference is given to elementary texts which cover the points needed. The book is intended for those of mathematical maturity equal to that of college students who have had one year of mathematics in college, including the theory of logarithms and elementary graphics. The later chapters of the book are not recommended for those whose general maturity is less than that of college juniors, but the earlier portions may prove suitable for those of less maturity.

Many of the author's friends, colleagues, and pupils have given him encouragement or assistance in sifting out what statistical concepts are important for presentation in an introduction of this type, and in moulding the material into presentable form. For all

Digitized by Google

such encouragement and assistance he is deeply grateful. In particular, he wishes to record his appreciation of the statistical ideas, encouragement to proceed with this project, or assistance of various sorts which he has received from four men who have been at one time or another his military, academic, or professional superior officers — Colonel Leonard P. Ayres, Colonel Malcolm C. Rorty, Professor Roland G. D. Richardson, and Dr. Edmond E. Lincoln.

When it comes to specifying what other writers in this field have been drawn upon most frequently or significantly, it seems impossible to be complete and accurate without devoting more time to historical research than is available. Certain statistical fundamentals are used so generally, though perhaps in slightly varying form, that the originator is lost to view, while other concepts equally simple and important have, if we may judge by the usual discussions, been overlooked altogether. We have mentioned in the text, as occasion arose, writers who have contributed something rather out of the beaten track, to whom the reader may wish to refer for more details. Special acknowledgment must be made of the indebtedness of the present writer to Professor Karl Pearson, whose researches and energetic campaigning made modern Statistics a reality. The author's inadequate mention of the other builders of statistical science will, he hopes, be pardoned by them, in view of the special aim of this book — to try to bring together the practical and the mathematical methods of approach to statistical analysis.

Digitized by Google

CONTENTS

I. The Statistical Approach — What do we Mean by Statistics? 1
II. Ratio or Percentage Analysis 16
III. The Combination of Ratios 42
IV. Frequency and Quantity Distributions . . . 53
V. Averages and their Use 80
VI. Index Numbers 108
VII. Measures of Dispersion 135
VIII. Regression or Trend Lines 155
IX. Analysis of Time Series 177
X. Two-way Frequency Distributions — Two-Variable Correlation 197
XI. Elementary Ideas in Regard to Sampling . . . 229
XII. The Binomial Expansion and the Normal Curve . 238
XIII. Probable Error 257
XIV. Three and Four Variable Correlation . . . 265
Bibliography 283
Appendix
I. Decimal Values of Common Fractions 288
II. Denominators of Coefficients in Parabolic Regression Equation 294
III. Ordinates and Areas of Certain Theoretical Frequency Curves 296
Index 299

Digitized by Google

Digitized by Google

INTRODUCTION TO THE MATHEMATICS OF STATISTICS

CHAPTER I

THE STATISTICAL APPROACH — WHAT DO WE MEAN BY STATISTICS?

Why use statistical methods? By statistical methods, we shall mean methods of summarizing, comparing, and analyzing quantitative data of a certain type. The exact characteristics of such data will become clearer as we proceed; we state here merely that the data of the social sciences are normally of this type, and that the data of the exact and natural sciences have some of the characteristics of this type, in so far as errors of measurement or extraneous influences interfere with exact observation.

We shall grasp more of the real significance of the statistical method by considering a typical problem which calls for statistical analysis than by attempting at this time further refinement of the foregoing definition. The point of view back of the statistician's tables and diagrams is so much different from that of the man in the street, that the first step for the embryo statistician is some understanding of *why* the statistican regards his methods as the best, rather than a microscopic analysis of just *what* are the characteristics of statistical data and statistical methods.

Suppose we consider, then, a question which, in the opinion of a statistician, calls for his methods. The street-car employees of a certain city are on strike for higher pay; should their demands be granted? The first answer to such a question which the ordinary man is likely to give is a vigorous assertion that the company (or the employees) is completely justified, and that the employees are a set of lazy loafers and trouble-makers (or the company is a hard-hearted machine for grinding the life-blood of the workers into dividends). The second answer, when it is found that such heat

Digitized by Google

does not produce light, is likely to emphasize particular acts of lawlessness or oppression for which one side or the other is responsible. But both methods of answering the question leave the unprejudiced bystander confused and unconvinced.

The first essential of any successful method of approach to such a problem is that it shall be scientific in temper, that is to say, that it shall attempt to discover and analyze all pertinent facts without heat, without prejudice, and without regard to which side they seem to favor. No satisfactory solution can be found by an investigator who decides on his conclusions first and then looks for facts to support them. The investigator may well say, of course, "From my previous knowledge and preliminary survey, I suggest that the following statements are probably true; I shall examine the facts to test this hypothesis." The formation of such hypotheses may be very fruitful in working up the material, if only an impartial attitude is still maintained, and the field of investigation is not unduly restricted to data bearing on these hypotheses.

In addition to an unprejudiced attitude of mind, however, a certain special type of thought is required. It is obvious that any judgment on the fairness of existing wages requires as a basis, among other things, a complete statement of what wages are being paid. This statement must be in a form which permits comparison with wages paid for similar work in that city or elsewhere. We shall find that the careful preparation of such a statement necessitates the use of certain processes of thought which are known as statistical methods. The same or other statistical methods are also found to be required in a thorough analysis of many other questions which arise in connection with the main problem of the fairness of existing wages. The statistician therefore regards his mode of attack as the only natural and sound procedure. He sees that most social questions can be solved only by adopting the scientific attitude, and that the scientific attitude, carried into such problems, leads inevitably to statistical methods.

"What wages are being paid?" The genesis of statistical series. Suppose that we are trying to state in definite form what wages are being paid. We may meet at once an apparent answer in

Digitized by Google

a statement that new conductors receive 60 cents an hour for the first six months, 65 cents an hour for the next six months, and 70 cents an hour thereafter; that motormen receive two cents an hour more in each service classification; and that operators of one-man cars receive 75 cents an hour for the first six months of such service, and 80 cents an hour thereafter.

In a way, this is a simple answer to our question, although it is not immediately convertible into weekly earnings, which must be known for fair comparison with other wages and with cost of living. But even this simple answer, when scrutinized, is found to contain certain fundamental statistical ideas. It has introduced order into a great mass of facts by using the principle of *classification*. All car employees are classified first on the basis of occupation, and second on the basis of term of service. If this process of classifying is explicitly recognized, it becomes natural to adopt the statistical device of *tabulation*, and to state the facts given above as follows:

HOURLY RATES OF PAY OF STREET-CAR EMPLOYEES IN *OUR CITY*

TERM OF SERVICE	OCCUPATION		
	CONDUCTOR	MOTORMAN	OPERATOR OF ONE-MAN CAR
First six months.....	60¢	62¢	75¢
Second six months.....	65¢	67¢	80¢
Twelve months or more..	70¢	72¢	80¢

There are several advantages in this method of arranging the figures, as compared with the textual method. In the first place, if we wish to pick out a single figure, it is much easier to find it in the table than in a series of sentences. For complicated data to which reference is to be made frequently, the tabular form is essential. It is unfortunately true that many readers have acquired a habit of omitting tables when they read, and for that reason, when we introduce a table in a discussion intended for the general public, it is perhaps wise to quote two or three figures from it in the text. The reader is then more likely to understand the table and to overcome his initial repugnance.

A second advantage of tabulation, besides facilitating reference, is that it makes the comparison of corresponding figures almost

Digitized by Google

automatic. Figures which are in the same horizontal line or the same vertical column, are brought into the field of vision together, and differences become striking. The mere arrangement of the figures in a table, then, goes far to accomplish one of the main aims of statistical methods — comparison.

The beginner should not assume that, because a table looks simple, he can comprehend it at a glance. It is important to acquire the habit of reading carefully the captions (i.e., phrases above the columns of figures) and stub (i.e., phrases to the left of the figures) as well as the title, notes, and any explanations that may be given. Title, captions, and stub in a sense give the framework in which all the figures of the table might be inserted to produce many sentences of text.

Let us look more closely at our table. The entries in the top row give the hourly rate for men of the same term of service in three different occupations. Such a series of figures is known as a *qualitative statistical series*, based in this case on occupational classification. Similarly the entries in the first column give the hourly rate for conductors of various terms of service. Such a series of figures is known as a *quantitative statistical series*, the basis of classification being the magnitude of some quantity, in this case, term of service.

The essential points in the construction of a statistical series are as follows:

(1) Two or more classes or groups are established, with exact enough boundaries so that every individual can be assigned definitely to some one class.

(2) We then either count the number of individuals who fall in each class, or note the magnitude of some measurable quantity which is typical of the individuals assigned to each class. The "measurable quantity," in the tabulation above, is hourly rate of pay.

In practically every question which involves a number of individuals, who differ from each other in several characteristics, it will be found that any scientific analysis requires the construction of statistical series, that is to say, grouping the individuals into classes and then either counting the number in each class or measuring

Digitized by Google

some characteristic of those in each class. As the methods of treatment which are most appropriate for statistical series differ according to the general type to which they belong, that is, according to the sort of classification which underlies them, it is worth while to note here the fundamental types.

Types of statistical series. We have met already one instance of the *qualitative* type of series, in which the basis of classification is occupation. Another basis of classification, of this general sort, is marital status — single, married, widowed, or divorced. It is obvious that this classification is important in any discussion of the adequacy of wages, as we should want to know the wages of married men as a group to compare with their needs. In any such comparison, it is clear that hourly rates of pay are not as important as actual amounts of money received in a week. The "measurable quantity," therefore, in a qualitative series based on the marital status classification would be, preferably, weekly earnings rather than hourly rates. The result of this classification would yield the following series:

Marital Status	Weekly Earnings
Single	$29.32
Married	34.68
Widowed	32.50
Divorced	37.10

In order to reduce to a single figure the divergent figures for weekly earnings of different single men (or men in any other group), it is necessary to use one or the other of the statistical means of summarizing which will be discussed more fully later. Let us suppose here that we adopted the familiar scheme of finding an *average*. We add the figures for weekly earnings for all the single men, and divide by the number of men. The result we regard at this point merely as a typical magnitude of the measurable quantity "weekly earnings" for the group of single men.

As soon as we attempt to study the fairness of the wages paid, we feel the need of making comparisons with other street-car systems in other cities. Suppose we find that in three other cities of about the same class, the hourly rate of pay of conductors of over twelve

Digitized by Google

months service is 65 cents, 68 cents, and 72 cents, respectively. These facts may well be listed as follows:

City	Hourly Rate of Pay
Grassville	65¢
Old Haven	68¢
Our City	*70¢*
Blue Falls	72¢

What we have obtained in this way is known as a *geographic series.* We have classified the individual conductors on the basis of location, and then observed a common characteristic of all individuals in each group.

The order of arrangement of the terms of this series, according to the magnitude of the observed or measured quantity, is one frequently used. A series so arranged is sometimes called a *rank list.* Another order is to move in some systematic way from one part of the country to another — say from east to west. Frequently, for convenience of reference, the *alphabetic* order is best.

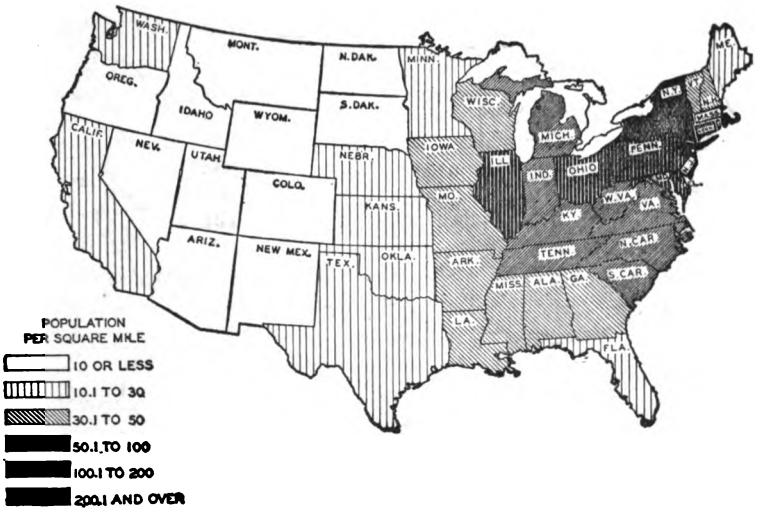
In view of the difficulty of reducing geographic series to the list form of arrangement, the results are often expressed diagrammatically by means of a map. Suppose, for instance, that we wish to present results which consist of a figure for every State in the United States — the density of population, let us say. Various methods of doing this may be found in the *Statistical Atlas of the United States.* Among the possible methods are the following:

(1) On an outline map, the figure for each State may be printed in that State.

(2) Five or six different types of shading may be used, all States in which the density is less than a certain figure being marked with the lightest shading, those in which the density is equal to or greater than the first limit, but less than a certain higher figure being marked with the next lightest shading, and so forth. Map No. 1 shows on this basis the density of population in the United States at the date of the 1920 Census.

Two other schemes are appropriate to picture *population* rather than *density* of population, namely:

Digitized by Google



MAP NO. 1. DENSITY OF POPULATION PER SQUARE MILE IN 1920 IN EACH STATE OF THE UNITED STATES

(3) A black dot, proportional in area to the population of the State, may be constructed within the outlined boundary of each State. This method has not proved satisfactory except where only rough results are to be shown or where the differences between States are enormous.

(4) A number of small dots of uniform size may be used, the number for each State being proportional to the population of that State. Map No. 2 shows by this device the distribution of the population of the United States in 1920.

The reason that Methods (3) and (4) would not be good for showing density is that some of the smaller States would require such large dots, or so many dots, that the dots could be located in or near the State only by selecting a scale which would make the dots in the larger States look insignificant.

Any discussion of wages is sure to lead to a comparison with what they were in earlier years. It may be, for instance, that two years ago conductors were paid only 64 cents an hour, where they are now paid 70 cents, and that back in 1914 the rate was only 32 cents.

Digitized by Google



Map No. 2. Population in 1920 of Each State of the United States
Each complete dot represents a population of approximately 500,000

If these facts are listed, with figures for intervening years, there results a full-fledged *historical series*, as follows:

Year	1914	1915	1916	1917	1918	1919	1920	1921	1922	1923	1924
Hourly rate for conductors	32¢	31¢	34¢	40¢	60¢	80¢	85¢	62¢	64¢	69¢	70¢

We have here classified the facts about individuals on the basis of *when* they were true, and observed a typical characteristic of all such facts in a particular year. It would almost invariably require a careful investigation to determine this "typical characteristic," as it would be concealed by a mass of apparently contradictory or at least divergent facts about particular individuals at different times in the year.

The series just stated might well be shown graphically, by a simple curve on a grid, the series of years being marked off at equal intervals along the horizontal base, the points checked directly above the years at heights proportional to the hourly rates, and consecutive points joined by straight lines. The result is shown in Diagram I.

We have already met a simple case of the quantitative series, in which the basis of classification is length of service. We can under-

Digitized by Google

stand other kinds of quantitative series somewhat better by thinking how we would analyze the weekly wage checks received by conductors. If the number of hours worked per week is somewhat irregular, these checks will not be uniform in amount. It will help



DIAGRAM I. HOURLY RATES OF WAGES FOR STREET-CAR CONDUCTORS, 1914–1924

us to understand conditions if we sort the checks into groups according to the size of check, putting in one pile all as large as $20, but not as large as $25; in a second pile all as large as $25, but not as large as $30; and so on. We may then count the checks in each group and add their amounts, and our results may be tabulated:

	CHECKS	AMOUNTS
$20 but under $25	10	223.56
25 but under 30	138	3,819.80
30 but under 35	2,428	80,612.60
35 but under 40	1,213	38,770.00
40 but under 45	116	4,820.00

The first column is known as a *frequency distribution*, and the second may be called a *quantity distribution*, although the latter term is not yet established usage.

The frequency distribution, while not in such common use as historical and geographic series, is really implied in thoughtful answers to questions of ordinary life. Suppose we are asked how

Digitized by Google

old college freshmen are; how many miles can a Ford run on a gallon of gasoline; how long does a student study his assigned lesson in mathematics. In each case the answer is, in some instances one thing, and in others another. Or more carefully, perhaps, we say, in answer to the first question, a few students are only 16; more are 17, a still larger number 18; while at the other end of the scale, there are a few scattering students over 22, with several 22, a larger number 21, and so on. If we want to state the facts somewhat more briefly, we use certain devices for summarizing frequency distributions which will be discussed in Chapter IV.

There is another type of quantity distribution, in which the individuals are classified on the basis of the magnitude of one quantity, and then the amounts of another quantity are added for all individuals that fall in each class. For instance, we might sort the weekly pay checks on the basis of the *age* of the workers, and then add the wages received for those in each age group. If we now write down the two columns used above, giving number of checks and total amount, it is obviously desirable also to put in a third column, obtained by dividing the entry in the amount column by the corresponding entry in the number column. We thus find the average wage for those of a given age. The table then will be as follows:

NUMBER AND AVERAGE WEEKLY EARNINGS OF STREET-CAR EMPLOYEES

AGE GROUP	NUMBER	AGGREGATE EARNINGS	AVERAGE EARNINGS
Under 25	130	4,102	31.55
25–44	2,600	87,100	33.50
45–64	1,100	35,310	32.10
65 and over	75	1,734	23.12

Unfortunately the above figures are fictitious; while a few concerns have tabulated the figures to show such facts, the results are not published in detail because of their confidential nature. More information is available as to the variation of the earnings of engineers, doctors, lawyers, and teachers with age than is available for wage earners.

The third column in this table is an example of a new type of

Digitized by Google

quantitative series, or rather, a somewhat different form of the simple quantitative series formed by the first column of the first table in this chapter. As each term in this series is the magnitude of a quantity (in this case earnings per employee) and this magnitude depends upon the magnitude of another quantity (in this case age), such a series may be called a "functional quantitative series." The word functional is used in the technical mathematical sense, that one variable is a function of another, if its value is definitely determined when a value of the second quantity is determined.

A functional quantitative series may be presented to the eye by a graph, of the same sort as used for historical series. As the series stated above is too concentrated to make much of a picture, we shall postpone actual illustration to Chapter VIII, in which historical and functional quantitative series will be discussed in detail.

An argument might be advanced for including "rank lists" among quantitative series. It seems preferable, however, to classify series according to the principle of classification rather than according to the principle of arrangement. The rank list arrangement is very common in dealing with qualitative or geographic series, and occasionally may be useful for a special purpose with historical series. Ordinarily, however, neither historical nor quantitative series are rearranged as rank lists.

Importance of the process of formation of series. The attempt, then, to arrange statistical data in a form which will assist forming judgments leads inevitably to a process of classification and then to the formation of statistical series. Such series are of four types — qualitative, geographic, historic, and quantitative. The formation of such series implies both arrangement and summarization of the crude data, and adapts the material, as will be seen hereafter, to the process of statistical analysis.

It is to be noted that each sorting ignores other differences in the given data than the ones which are basic in that particular classification. When we are finding how many checks fall between $20 and $25, $25 and $30, and so on, we do not consider the age of each individual. When we are finding how many checks went to men under 25 years of age, how many to men 25–44, and so on, we do not

Digitized by Google

consider also the classification by amount or by country in which father was born or other possible principles of differentiation. It is possible to use simultaneously a number of different bases of classification. The work, however, becomes increasingly complicated to carry through, and the results of too fine a division are so difficult to grasp as to contradict the first aim of statistical analysis — the reduction of a complicated mass of quantitative data to a form easy to comprehend.

In our subsequent discussions, we shall find statistical series fundamental. Some chapters, for instance, will take up methods for discussing frequency distributions; others, methods for discussing historical series. Some writers have ascribed the great apparent difference between biometric statistics and economic statistics to the great use made of frequency distributions in the former, and the great use made of historical series in the latter. While we believe this difference has been overemphasized, there is no question that different methods are appropriate to the two series, and that a first step in statistical knowledge is the gaining of a clear idea of just what these two types of series are.

Exact definition of boundaries of classes necessary. In all sortings, definite criteria are necessary to determine, uniformly and without question, in which group a given item should be included. The census enumerator has to do a great deal of sorting, as he records the data regarding any individual. He must list each individual as Caucasian, negro, or other colored, as native or foreign born, must specify age, marital condition, country of birth, and other characteristics. Most people can be classified without great difficulty, but there are some doubtful cases, which can be resolved only by careful reading of instructions. Should a mulatto be recorded as Negro or as other colored? Should a person born in territory once German but now French be classified as French or German by birth? Is a person born on the high seas, in a vessel flying the American flag, native or foreign born? Shall age be recorded as of latest or nearest birthday?

The drawing up of instructions to cover such questions is not a topic in mathematical statistics. It is fundamental to any mathe-

Digitized by Google

matical methods, however, that the material shall have been classified consistently, and that the user shall know and use the same definitions of terms as were used in classifying the items. Figures from different sources are frequently compared without due allowance for the fact that the definitions used in one source differ from those used in the other. Much of the effort of a statistician who uses figures compiled by more than one agency should be devoted to discovering such differences in definition and if possible adjusting the figures to the same basis before combining them. If such adjustment has not been adequately done, it is worse than useless to apply elaborate processes of mathematical analysis.

Background statistics. The purpose of statistical analysis, as we have thought of it so far, has been to coördinate and arrange masses of quantitative facts as a preliminary to deciding some specific question. There is, however, a variant form of statistical work that at first sight does not seem to originate in the desire to secure light on a particular problem. But closer examination of the circumstances often shows that the same material really serves to illuminate many problems, and that the need of such material is so generally recognized that no specific questions are explicitly stated. In fact the formulation of such questions in definite form may be difficult or impossible before the preliminary results of a statistical inquiry are available. Statistics of this sort are sometimes called "background statistics." While there is a danger that statistics resembling these may come to be an end in themselves, it is desirable to devote a good deal of attention to statistical analysis whose aim is a description of existing conditions rather than problem solving.

A good instance of valuable background statistics is the United States census. The results serve to describe conditions as they were at a specific time. With these results in hand, it is possible to formulate specific inquiries as to what can be done to improve conditions, or to plan a statistical investigation as to whether a particular business concern is getting its due share of the sales of its product in every State.

Another instance of this general sort of statistics is the current

Digitized by Google

reports of a railroad or manufacturing corporation. Such reports in themselves do not solve any problems of management, but they often tell the executives whether something is wrong or not, and perhaps take the first steps in locating the difficulty. Special investigations, often also of a statistical nature, are necessary as a guide to definite executive action.

Statistical coöperation. It is clear from preceding sections that statistical investigations require great care in details as well as a big bulk of work. It is obviously impossible for one worker to analyze both the problems of his particular business and related problems in other lines, or both his scientific problem and other related problems, if he tries to do all the detailed analysis himself. Coöperation of all is essential to the success of each. Such cooperation implies, among other things:

(1) A frank statement of each worker as to the assumptions made and the sources of his information, as well as complete honesty both in work and presentation;
(2) As far as practicable, uniformity in method and use of terms;
(3) Full statement of definitions and of bases of classification.

Essential points in the statistical approach. The essential points in the line of thought which leads to statistical methods, as outlined in this chapter, may be summarized as follows:

(1) Complicated social questions require that the attitude of the investigator shall be *scientific*, that he shall be free from prejudice and willing to examine all pertinent facts in detail.
(2) When such an investigator begins to examine the facts bearing on any problem, he finds them so numerous and complicated that he is compelled to *classify* them.
(3) Such classification leads to the construction of *statistical series*.
(4) Such series may themselves be classified, according to the basis of classification, as *qualitative*, *geographic*, *historical*, or *quantitative*.
(5) Such series constitute the raw material for the application of processes of *statistical analysis*.

Digitized by Google

(6) Such statistical analysis will be useless if the original classification is carelessly done.

(7) Statistical analysis may have as its aim either the guidance of those who have to decide specific problems, or the description of the *background of fact* behind specific problems.

(8) Because of the difficulties in detail and the extensive character of statistical investigations, coöperation between workers is necessary. This means honesty, uniformity, and completeness.

EXERCISES AND PROBLEMS

Exercise 1. Tabulate how many *a*'s, *e*'s, *i*'s, *o*'s, and *u*'s there are in one page of some book.

Exercise 2. Consult the *Statistical Abstract of the United States*, or the *World Almanac and Book of Facts*, and note the statistical series there shown. Classify each such series under the types outlined in this chapter; decide whether the titles, captions, stubs, and accompanying notes are adequate to enable the reader to understand the basis of classification, and real meaning of the figures; note tables which show classification on two, three, or four bases simultaneously.

Problem. Select some problem of present and preferably local interest. Try to outline the statistical material which would be of value to those who have to deal with this problem. As far as available data permit, draw up statistical tables presenting this material, with necessary explanation and such obvious graphical illustration as seems appropriate. Do not try to carry the investigation to a point which requires more technical knowledge of statistics than you have as yet, but note difficulties which arise and press the investigation further as statistical methods are explained in succeeding chapters.

The following problems are suggested:

(1) Should a new high school building be built?

(2) Should the number of students in the University be limited, and if so, to what number?

(3) Is this city growing as rapidly as similar cities in population, bank clearings, school attendance, births, trains per day, and other measures of progress?

(4) Should the street railroad be permitted to raise its fare?

(5) What sort of men are listed in *Who's Who in America?* Select a group of such names by some arbitrary process, such as taking the second name on each page, from page 200 to page 500, and note their ages. Their education, occupation, and other characteristics may also be summarized, and such two- or three-fold classifications made as seem desirable.

Digitized by Google

CHAPTER II

RATIO OR PERCENTAGE ANALYSIS

AFTER statistical series have been constructed, the relationships between terms in a series, between one term and the whole group, and between corresponding terms in parallel series, are made clearer by the use of *ratios*, often in the form of *percentages*. The use of percentages is so familiar that we often fail to realize its significance. An investigator should not find the ratio of a to b unless he thinks that the meaning of the facts can be brought out more clearly in that way. He should not include a ratio or percentage in his report unless in his opinion it actually does help the reader to understand the facts. Stating a percentage in a report therefore implies a judgment that it is significant, and finding a percentage is a step in the *analysis* of a statistical series.

In the following sections are discussed various ways in which finding the ratio of a to b may bring out the meaning of the facts.

Partition percentages. The first and in a way the simplest use of percentage analysis is to express the ratio between one term in a series and the sum of all terms. Thus we may express the number in each age group as a percentage of the whole population, or the number in each occupation as a percentage of those gainfully employed, or the population of each State as a percentage of the population of the entire country, or in general the terms of any qualitative or geographic series, or of quantitative series of the frequency or quantity distribution type, as percentages of the sum of the series. Historical series are seldom studied in this way, except that the sales or earnings for each month or quarter are often expressed as percentages of the sales or earnings for the year.

The following examples will show how the finding of these ratios brings out the real meaning of the facts, and also how such facts may be presented.

Digitized by Google

AUTHORIZED ENLISTED STRENGTH OF THE REGULAR ARMY OF THE UNITED STATES

BRANCH OF SERVICE	MEN	PER CENT OF TOTAL
Infantry	48,849	36.7
Cavalry	10,580	8.0
Field Artillery	17,736	13.3
Coast Artillery Corps	14,420	10.9
Air Service	8,760	6.6
Medical Department	7,235	5.4
All Others	25,373	19.1
Total	132,953	100.0

UNITED STATES POPULATION BY AGE GROUPS IN 1920

AGE PERIOD	PERSONS	PER CENT OF TOTAL
Under 5 years	11,573,230	10.9
5 to 14 years	22,039,212	20.9
15 to 24 years	18,707,577	17.7
25 to 44 years	31,278,522	29.6
45 to 64 years	17,030,165	16.1
65 years and over	4,933,215	4.7
Age unknown	148,699	0.1
All ages	105,710,620	100.0

POPULATION OF NEW YORK CITY, BY BOROUGHS, IN 1920

BOROUGH	PERSONS	PER CENT OF TOTAL
Manhattan	2,284,103	40.6
Brooklyn	2,018,356	35.9
Bronx	732,016	13.0
Queens	469,042	8.4
Richmond	116,531	2.1
All boroughs	5,620,048	100.0

TAX LEVIED BY UNITED STATES ON NET INCOME OF CORPORATIONS

(Income earned in calendar year 1921)

INCOME CLASS	TAX	PER CENT OF TOTAL
$2,000 to $5,000	$4,917,588	.7
5,000 to 10,000	12,562,214	1.8
10,000 to 50,000	79,327,566	11.3
50,000 to 100,000	56,238,078	8.0
100,000 to 250,000	86,935,248	12.4
250,000 to 500,000	69,398,858	9.9
500,000 to 1,000,000	68,819,417	9.8
1,000,000 to 5,000,000	162,226,494	23.1
5,000,000 and over	161,149,969	23.0
All classes	701,575,432	100.0

Digitized by Google

Advantages of the use of percentages. Before we go on to discuss other somewhat different uses of ratio analysis, let us pause to consider just what we have gained when we put ratios in the form of percentages, as we have done in the examples just given. The main advantage is the greater ease with which we can make comparisons. Frequent use of the percentage scheme has made us all thoroughly familiar with the idea of dividing a unit into a hundred parts. Instead of having mental pictures of all possible ratios, we carry in our minds one clear picture of the relation to 100 of all integers less than 100, and refer all other ratios to that scale. This mental concept is of course closely related to the decimal system of notation, and even more closely to our system of coinage. Such and such a per cent of total expenditures or receipts is, as we often say, so many cents on the dollar. Our unfortunate British cousins have to replace this simple idea by one of so many shillings and pence in the pound, that is to say, by division into twentieths and then into twelfths.

As noted in the first paragraph of this chapter, finding a ratio between two quantities is a step in the analysis; putting this ratio in the form of a percentage is not further analysis, but merely a device to make the facts easier to assimilate.

Computation of percentages. As successful statistical work depends in considerable measure on establishing expeditious methods for carrying through the routine processes which are found to be desirable, it is worth while here to discuss briefly the methods of computation of percentages.

The arithmetic process involved in finding the ratio of a to b as a decimal fraction is merely division; to express as a percentage, we merely move the decimal point two places to the right in the quotient. If inefficient methods of division are employed, this operation, when required in bulk as it often is in statistical work, will consume an excessive amount of time and energy. A worse consequence is that the results may be honeycombed with errors.

The possible methods are as follows:

(1) Computing machines.
(2) Slide rules.

Digitized by Google

(3) Logarithms.
(4) Multiplication or division tables.
(5) Long hand division.

For any considerable amount of work for which accuracy is desired, the use of computing machines is very much superior to any other method. The selection from among the many satisfactory machines now available should be guided in part by the way in which the machine is to be used. If the statistical working group includes several persons, the purely routine computation may be sufficient in amount to justify the assignment of one or more persons exclusively to that work. In such cases the machine should be selected with which the greatest volume of work can be turned out accurate to the required number of significant figures, even if some special training of the operator is required. If only three or four figures are required in the results of the divisions or multiplications, satisfactory results are usually secured from one of the machines which perform multiplication by repeated addition and division by repeated subtraction. If greater accuracy is needed, it may be desirable to secure one of the somewhat more expensive machines which really perform multiplication as a single operation.

In cases in which actual computation occupies only a part of the worker's time, maximum speed is a less important consideration in the choice of a machine than simplicity of operation. A machine on which the four fundamental operations can be performed accurately by an intelligent person after five or ten minutes instruction, is superior in such cases to a machine which requires special training. It is recommended that a machine of the former type be available even when one worker is assigned exclusively to computing, using a machine requiring special training. It often happens that a few results are needed at once, in the course of a statistical investigation, and time and momentum would be lost by delaying until the computer could be diverted from work in hand.

If division is performed by machine, the location of the decimal point in the quotient should be carefully checked by some one familiar with the question under consideration, as machine operators often make mistakes in pointing off.

While a computing machine is to be preferred for any extensive

Digitized by Google

computations, it often happens that a statistician does not have immediate access to a computing machine, and is willing to allow errors of one part in a thousand in his results. In such cases, the use of the logarithmic slide rule is to be recommended. Some computations, moreover, can be done more rapidly on the slide rule than on the computing machine; for instance, multiplication or division of a series of numbers by the same number. Any statistician will find it worth while to be able to use the slide rule for multiplication, division, and extraction of square root, and if already familiar with logarithms, should be able to acquire enough understanding of the instrument from a few hours study of the instruction sheets to begin to make use of it.

Instances sometimes arise when somewhat greater accuracy is desired than is guaranteed by the slide rule, but a computing machine is not available. In such cases, the use of four place logarithms may be advantageous. The great advantage of a four place table, as compared with a five, six, or seven place table, is that no turning of pages is required. The time consumed by one thoroughly conversant with logarithmic computation will not be much greater than with a machine, and the error in the results will seldom exceed one part in five thousand. If greater accuracy is desired, the larger tables must be used, and the time required will be considerably greater. In such cases greater effort to get at a machine is recommended.

Multiplication or division tables have the advantage of absolute accuracy, but are ordinarily unsatisfactory substitutes for a machine. It is believed, however, that the table in the Appendix of this book will occasionally be useful, especially to those who are working where a machine is not available. By this table all ratios in which the denominator is a two digit number may be reduced to percentages by simple addition. The table gives the succession of digits in the quotient when each of the numbers 1, 2, 3, 4, 5, 6, 7, 8, 9 is divided by each of the numbers from 11 to 101.

While long hand division is slow and subject to inaccuracies, it is sometimes a saving in time to do one or two divisions in that way rather than to stop work to get an assistant to do them or to walk some distance to a machine. If only the first two or three figures

Digitized by Google

are wanted, but wanted without possibility of doubt, long hand division is sometimes better than slide rule or machine computation, if the latter involves delay.

A word should be said in regard to the desirability of absolute accuracy in computations which are to be published. Careful statisticians do not expose themselves to unfriendly criticism by publishing figures which are strictly speaking one unit wrong in the last place. If therefore there is any doubt about the last digit after a slide rule computation, some other method should be used to see which way it falls. For instance, suppose it is found that 106.5 seems to be correct, and only three digits are to be retained. Shall the number be called 106 or 107? A machine computation may show that the number was really 106.48 or 106.53; in either case the slide rule would give 106.5 and apparently not decide the question. It is true that it is not a matter of great importance in most statistical work, as far as actual significance goes, whether 106 or 107 is used, but an impression of carelessness is conveyed to any reader who notices that 107 is used for a number which is really 106.48.

A slight exception may be noted. When we are finding what per cent each part is of the whole, it is desirable that the total of these per cents shall be exactly 100, even though the ordinary rules for rounding off lead to 99.98 or 100.03. In such cases, it is well to carry the computation one place further than it is desired to retain, and then round off in such a way as to do the least violence to the face of the returns and yet add up to an even 100.

The preceding paragraph will perhaps make it clearer why machines are so much preferred to other methods. With them it is possible to secure unquestioned accuracy without as much expenditure of time as equally accurate work by other methods requires. Some of this accuracy is not really significant, perhaps, but it contributes to making a really workmanlike job.

Graphic treatment of partition percentages. The division of a whole into parts may be shown graphically by any one of the following devices:

(1) A bar divided into parts, each with a different shading;

Digitized by Google

(2) A circle with sectors proportional in area to the different parts (known as the *pie diagram*);

(3) A series of parallel bars or columns. This scheme emphasizes the comparison of the parts with each other.

These three methods are illustrated by the diagrams shown herewith. The *Statistical Atlas of the United States*, the *Monthly Bulletin* of the New York Federal Reserve Bank, and other statistical publications contain many diagrams constructed on the same principles.

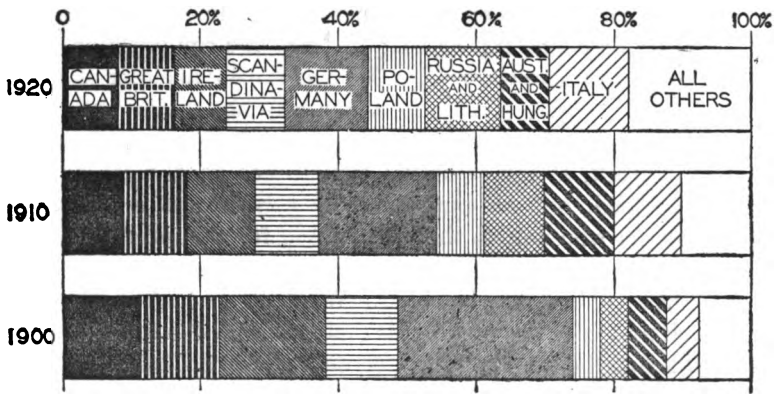


DIAGRAM II. PERCENTAGE OF THE FOREIGN-BORN INHABITANTS OF THE UNITED STATES WHO WERE BORN IN EACH OF THE COUNTRIES INDICATED

Each bar in this diagram illustrates the first method for presenting the division of a whole into its parts; taken together they compare the results of such division for three censuses.

Instead of the self-keying method here used, statisticians frequently give at one side or the top or the bottom of the diagram a "key" showing what each color or shading means. Where it is practicable, the self-keying method is to be preferred, as it does not require diversion of attention from the main diagram. The main point of graphic methods is to put the material in a form which can be understood quickly; any feature of a diagram which delays its comprehension by the reader is therefore in so far objectionable.

If it is desired to stress the comparison of the percentages born

Digitized by Google

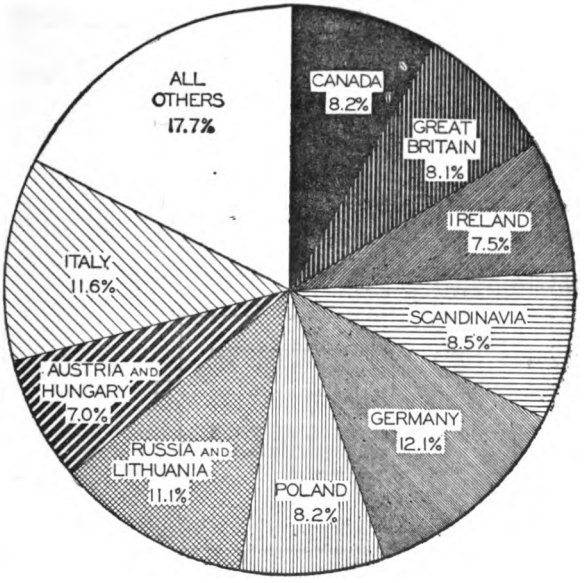


DIAGRAM III. "PIE DIAGRAM." PERCENTAGE OF THE FOREIGN-BORN INHABITANTS OF THE UNITED STATES IN 1920 WHO WERE BORN IN EACH OF THE COUNTRIES INDICATED

in one of these countries as of the various census dates, it has been found that the use of three "pie" diagrams, one for each date, is unsatisfactory. The eye has difficulty in picking out corresponding sectors definitely enough so as to grasp the differences. The selection of corresponding portions of bars in Diagram II is also unsatisfactory except for the two countries shown at the ends. It is usually found preferable, if the emphasis is on the comparison of percentages at the three dates, to use the type of chart shown in Diagram IV.

This diagram enables one to compare the percentages for the various countries according to the 1920 (or 1910, or 1900) Censuses, but does not facilitate that comparison so much as if the bars for any one Census were grouped together. This has been done in Diagram V.

Diagram V is perhaps the best way of showing the facts for gen-

Digitized by Google

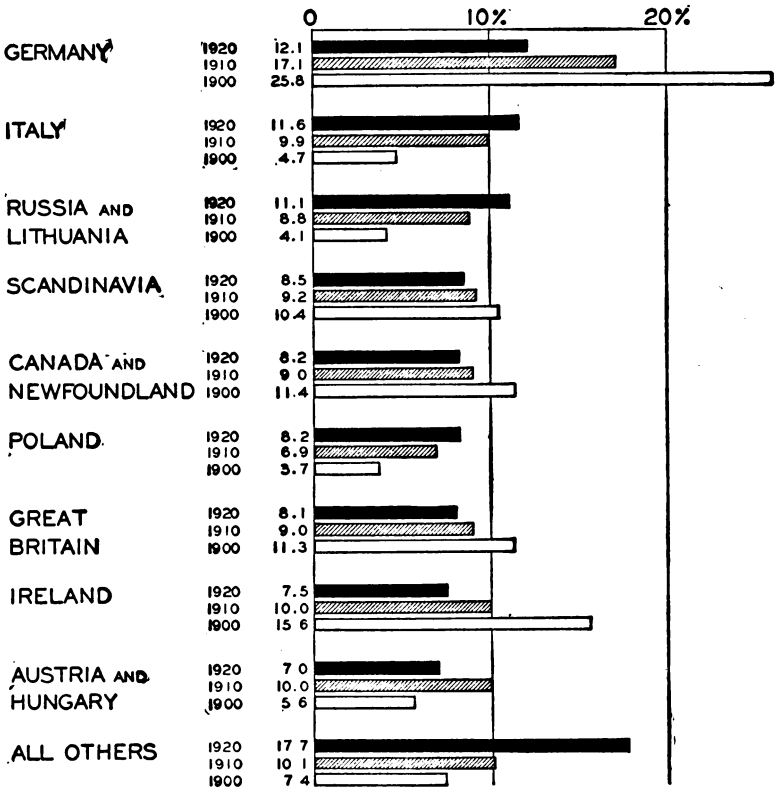


DIAGRAM IV. PERCENTAGE OF THE FOREIGN-BORN INHABITANTS OF THE UNITED STATES WHO WERE BORN IN EACH OF THE COUNTRIES INDICATED, ACCORDING TO THE 1920, 1910, AND 1900 CENSUSES

eral purposes, although Diagram IV is somewhat more compact and thoroughly satisfactory if only two Censuses are to be compared. With three sets of bars, it begins to be confusing, and with four sets, hardly carries out the fundamental purpose of a graphic device — to present the facts in such a way as to be more easily grasped and remembered.

Use of percentage to measure progress relative to a standard. When the partition of several groups is made on the same basis, and

Digitized by Google

the results of each partition expressed by percentages, a comparison of similar items can be made at once. Diagram II taken as a whole is an instance of this. If the partition is merely twofold, the percentage in one class is merely 100 minus the percentage in the other, and therefore only one percentage needs to be stated. A compari-

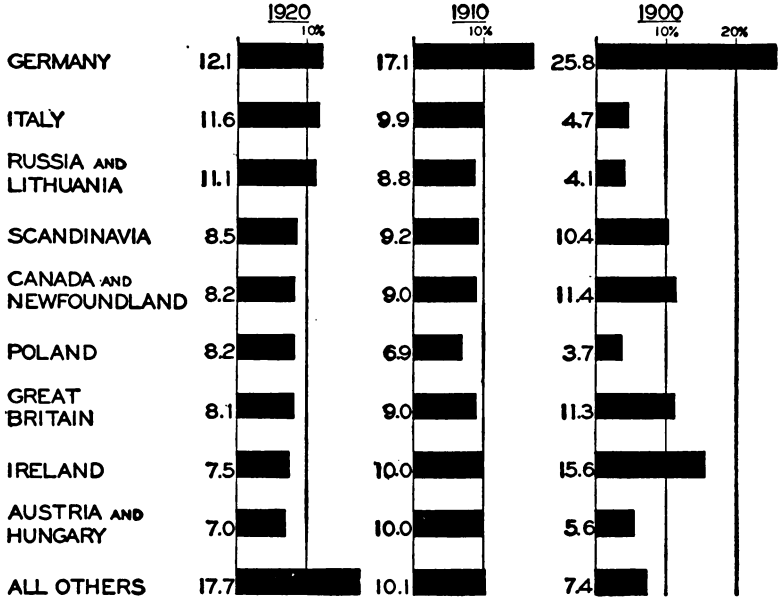


DIAGRAM V. PERCENTAGE DISTRIBUTION OF THE FOREIGN-BORN INHABITANTS OF THE UNITED STATES BY COUNTRY OF NATIVITY, AT THE 1920, 1910, AND 1900 CENSUSES

son of these percentages for the different groups tells the whole story, and the table or graph may be reduced to a column of these percentages or to a series of bars proportional to those percentages. A simple instance of such a table is the daily report of the standing of major league baseball teams in the newspaper, where the "percentage" [really "permillage," as 1000 is the scale of reference, not 100] shows for the games played by each club the percentage won. From this usage we pass readily to what is really something different. When the division is twofold, the percentage won, or similar

Digitized by Google

figure, is really a measure of performance related to a standard. The ratio, $\frac{\text{actual}}{\text{standard}}$, therefore becomes a significant figure, best expressed as a percentage. We have, for instance, government crop reports expressing the condition of the crops as a percentage of an ideal "normal." Or a business concern may express its current progress by the percentage the sales of each department are of some assumed quota, or "budget."

During the World War, the United States War Department was confronted with the problem of securing enough rifles, blankets, airplanes, etc. to equip and maintain the troops. In order to measure how successfully this was being done, the statisticians of the Department make considerable use of exactly this type of diagram, as shown in Diagram VI herewith. For any given date, the amount or number of each item which should have been made available in order to meet all needs, was determined, and the amount which had actually been secured was expressed as a per cent of these requirements. A glance at such a diagram made it possible to select those items which from this point of view needed most attention.

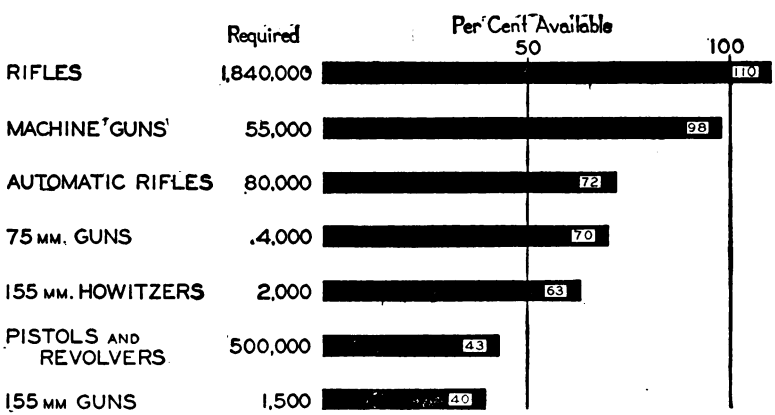


DIAGRAM VI. PERCENTAGE OF REQUIRED SMALL ARMS AND ARTILLERY AVAILABLE JULY 1, 1918
Figures fictitious, for illustration merely

Digitized by Google

Interest rate, death rate, and birth rate. In certain cases, it has become customary to use as a measure the ratio of two related numbers, even when there is no implication that the standard of full accomplishment is 100 per cent. Three well-known ratios of this sort are interest rate, death rate, and birth rate.

The interest rate is the ratio of the sum paid for the use of a sum of money for a year to the sum lent. An interest rate is practically never larger than 20 or 25 per cent. Everybody, especially bankers and business men, has a background of knowledge in regard to the rates of interest which are paid or have been paid under various conditions. If we are writing for an audience which does not have special knowledge of interest rates under the conditions we are discussing, it is desirable to furnish enough information to serve as a basis of comparison. We cannot assume that comparison can be automatically made with an obvious standard, as we could when dealing with "per cent of requirements available" or similar problems.

The annual death rate is the number of deaths in a year divided by the total population at the beginning of the year, the result being put on a "per thousand" basis by multiplication by 1000. The number of deaths in a year is often, for the purposes of this computation, not the actual result of a full year's experience, but the number that would die in a year if a shorter period — e.g., week or month — may be considered typical. In other words, the annual death rate is given by the formula:

$$\text{death rate} = \frac{1000 \text{ times actual deaths times number of periods in year}}{\text{Total population at beginning of period}}$$

For instance, if 15 people in a group of 30,000 die in a certain month, the annual death rate for that group of people for that month is

$$\frac{1000 \text{ times } 15 \text{ times } 12}{30{,}000} = 6$$

Any particular value of the interest rate, death rate, birth rate, or other similar rates, is significant only to those with adequate information with regard to the rate in question. No unfamiliar rate should be used or referred to without full discussion and exem-

Digitized by Google

plification, unless 100 per cent may properly be considered standard. If 100 per cent is not actually the comparison which one has in mind in using the rate, an attempt should be made to determine what is the standard which those who use the rate really have in mind. For instance, vital statisticians know that the death rate for the entire population in the United States registration area has been running between 11 and 15 for the past 15 years, except for the one year 1918, when it rose to 18. They know that these rates are lower than in earlier years. They know also that the death rate is lower for whites than for negroes, that it varies with age, being smallest for children 10 to 15 years of age and very high both in the early years of life and above the biblical three score years and ten. It may almost be said that the value of the judgment of a vital statistician as to the significance of a new development is as much affected by his background of digested information of this type as by his native ability.

Another instance of a ratio which has great significance for those with the necessary background is the "operating ratio" — operating expenses divided by total income — as used by railroad statisticians. Other instances of ratios frequently used are as follows:

"Labor turnover" — Number leaving per year divided by average number employed.

"Current ratio" — Ratio of current assets to current liabilities.

"Tax rate" — Ratio of tax to assessed valuation.

When a student meets any rate of this type for the first time, he should try to gather and remember enough typical values the rate has had under different conditions, so that his judgment in regard to the next use of it he meets may be based on fair comparisons.

Instance of an inadequately explained rate. The desirability of explaining a rate which is not to be compared with 100 per cent as a standard will be brought out by considering an instance in which no explanation is given. In the *Reports of the United States Bureau of Education*, the "per cent of students graduated" from secondary schools is shown without any comment to indicate that anything other than 100 per cent is to be taken as the standard with which to

Digitized by Google

make comparisons. On hasty reading, therefore, one gathers the impression that a ratio of 20 per cent is very poor. A little reflection enables us to realize that 25 per cent is a sort of ideal standard, which would be attained only if every one who entered were graduated four years later and all classes were of the same size. If, as is usually the case, the more recent classes are larger, perfection would be indicated by a figure considerably less than 25 per cent. In view of the fact that no one can judge what a fair standard is for the ratio of graduates to students, unless a good deal more information is provided, a better measure to use in this discussion would be the ratio of number of graduates to those who entered four years earlier.

Ratios which are not percentages. Standard ratios need not necessarily be expressed as percentages. In financial statistics other ratios are often used, such as cost per ton-mile (railroads), cubic feet per ton (shipping), cost per mile (railroad construction), bushels per acre (agriculture), production per employee (manufacturing), and many others. These ratios have the appearance of averages, and are often referred to as averages. In essence, however, they are primarily "size-ratios," as Zizek calls them, and do not attempt to select from a large mass of particular items a single one which may be considered typical. The statement that the yield of wheat in Ohio was thirty bushels per acre in a certain year does not mean that the figures for the first acre, the second acre, the third acre, and so forth were all found, and then their average computed, but merely that the total yield was divided by the total acreage. As we shall discuss in detail in later chapters, an average which has behind it detail of the individual items is of rather different significance than one which is derived as the quotient of two totals. In the latter case, it can not be analyzed further if such analysis is suggested by related facts.

Some ratios of this type which are in frequent use are merely standards by agreement, definition, or law — for instance, wages per hour in a particular trade in a particular city at a particular time, rifles per regiment according to a particular plan of army organization, number of inches in a meter, or number of pounds of coal in a ton. Such ratios are often important in statistical compu-

Digitized by Google

tations and in such use are treated in the same way as ratios derived by division of one observed statistical quantity by another.

No matter what problem a statistician takes up, he finds the use of such statistical ratios as above described is essential to success. In his preliminary analysis, he tries to discover ratios which are fairly constant, or which change only when there is a real difference in underlying conditions. If we are trying to compare different colleges and universities, for instance, we find that the number of students in a college varies from 42 to over 10,000, and the size of faculty from 8 to 1200. But the number of students per teacher is much more stable, and variations from the normal value of 12 to 15 are likely to indicate that the institution differs somewhat from the usual type or is under or over staffed.

When a student takes up a subject new to him, but already studied by others, he should give careful attention to the ratios already in use. He should note under what conditions the ratio remains about constant, what are considered typical values under various conditions, and how far values must differ from such typical values to indicate significant differences. Suppose he takes up educational statistics, for instance. He finds that great use is made of what is known as the "Intelligence Quotient" or "I.Q.," the ratio of mental age, as determined by psychological test, to chronological age. Experience has shown that the I.Q. of the same child, tested in different years, remains substantially the same. Experience has also shown about as many children will test above 100 as test below, and that a difference of 10 in I.Q. percentage indicates a real difference in mental ability. This outline of the significance of the I.Q. would of course have to be made more complete before the student would be ready to use the ratio with full confidence, but indicates the sort of information which is necessary when one takes up a new ratio.

In all work with ratios, except the most elementary, it is essential to understand clearly the difficulties which arise when it is desired to find the ratio which applies to a group, given the ratios which apply to individual items or to sub-groups. These questions will be discussed in the next chapter, under the heading, "The Combination of Ratios."

Digitized by Google

Relative numbers. The discussions of this chapter up to this point apply in general to series of all types. In the treatment of historical series, in particular, the ratio which is most useful is the ratio of each item in the series to the first item or to some other item which is chosen as the base. Suppose for instance that the price of a certain commodity on the first day of four successive months was 50 cents, 60 cents, 55 cents, and 65 cents. The ratios of these successive values to the first value, expressed as percentages, are clearly 100, 120, 110, 130. The numbers thus found are called relative numbers, and the process just performed is referred to as "reducing the given series to a series of relative numbers with the first year as base."

The advantages of this process are as follows:

(1) The meaning of the series is easier to grasp, for the same reasons that we discussed in connection with the simplest uses of percentages.

(2) Comparison of two or more series is facilitated.

(3) As we shall see in a later chapter, the series is prepared for further analysis, and for combination with other series.

This scheme is especially valuable in comparing the changes in two varying quantities whose absolute magnitudes differ widely. As an example of this, consider Diagram VII, which compares the growth of Great Britain, France, the United States, and Canada. The figures are as follows:

Year	Population in Thousands				Relative Numbers 1800 = Base Year			
	United States	Great Britain	France	Canada	United States	Great Britain	France	Canada
1800	5,308	10,501	26,931	240	100	100	100	100
1810	7,240	11,970			136	120		
1820	9,638	14,092	29,871		182	134	111	
1825				582				242
1830	12,866	16,261			242	155		
1840	17,069	18,534	33,401		322	177	124	
1850	23,192	20,816		1,842	437	198		768
1860	31,443	23,129	35,845	3,091	592	220	133	1,288
1870	38,558	26,072	36,103	3,689	726	248	134	1,537
1880	50,156	29,710	37,672	4,325	945	283	140	1,802
1890	62,948	33,028	38,343	4,833	1,186	315	142	2,014
1900	75,995	37,000	38,962	5,371	1,432	362	145	2,238
1910	91,972	40,831	39,605	7,207	1,733	389	147	3,003
1920	105,711	42,768	39,210	8,788	1,991	407	146	3,662

Digitized by Google

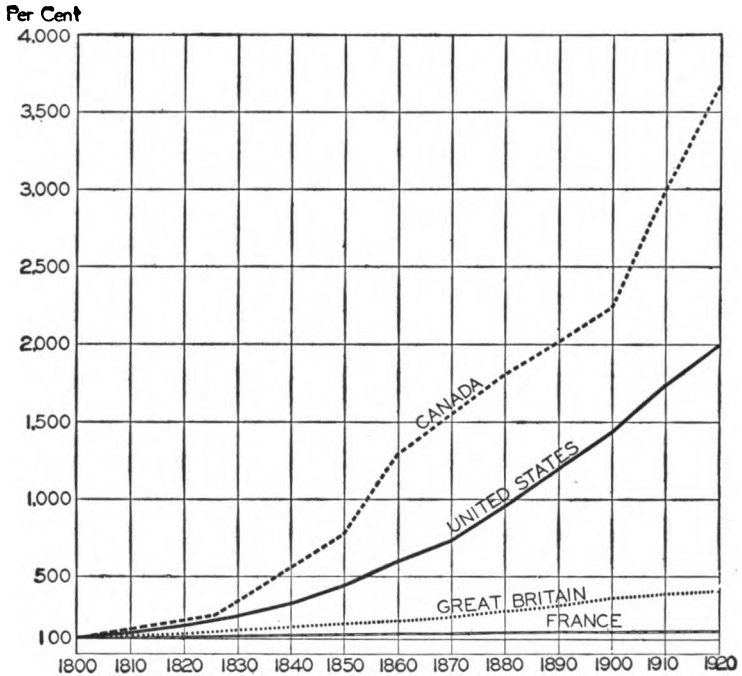


DIAGRAM VII. POPULATION OF GREAT BRITAIN, FRANCE, CANADA, AND THE UNITED STATES, REFERRED IN EACH CASE TO THE 1800 POPULATION AS BASE
French Figures exclude Alsace-Lorraine throughout

Relative numbers are closely related to *index numbers*, which will be discussed in detail in Chapter VI. Some writers in fact use the term *index number* for what we have here defined as relative numbers. It seems preferable, however, to separate this comparatively simple concept from the complications which inevitably arise in the discussion of true index numbers. The aim of index numbers is to summarize changes from time to time or from place to place in a large number of related quantities, and relative numbers are often used in the details of the analysis or in stating the results. As some index numbers, however, make no use of relative numbers, and relative numbers may be used in some problems where no notion

Digitized by Google

of index number is involved, the two concepts should be kept distinct.

The ratio chart, or semi-logarithmic diagram. If two historical series which differ considerably in absolute magnitude are to be compared, direct plotting on the ordinary uniform scale does not facilitate certain important comparisons as well as either one of two other methods. One of these methods is to reduce each series to relative numbers with the same year as base, and to plot as in Diagram VII, discussed above. The second method is less obvious and in a sense not so simple, but has distinct advantages, especially for series in each of which consecutive terms differ widely, and in which it is desired to compare changes rather than absolute magnitudes.

The machinery of this method consists of graph paper in which the vertical ruling is based on the logarithms of the numbers rather than the numbers themselves. The effect of this device is that equal vertical distances represent equal percentages of change. This principle may be stated more fully as follows:

If the vertical distance between two points on the chart is the same as the vertical distance between another pair of points, the ratio of the number corresponding to the upper point to the number corresponding to the lower point is the same for the first pair of points as for the second pair.

If each term of a statistical series is 20 per cent greater than the preceding, or in other words, the ratio of each term to the preceding is 1.2, the graph of the series on a chart of this type is a straight line, as the vertical rise from point to point is constant.

The advantages of this type of chart will be more evident after examination of Diagram VIII, which is based on the same material as Diagram VII.

The following points should be noted:

(1) By inspection or application of a straight edge, we see that the population of the United States from 1800 to 1860 (or the population of Great Britain from 1830 to 1910), increased during each decade by about the same percentage of the population at the beginning of the decade.

Digitized by Google

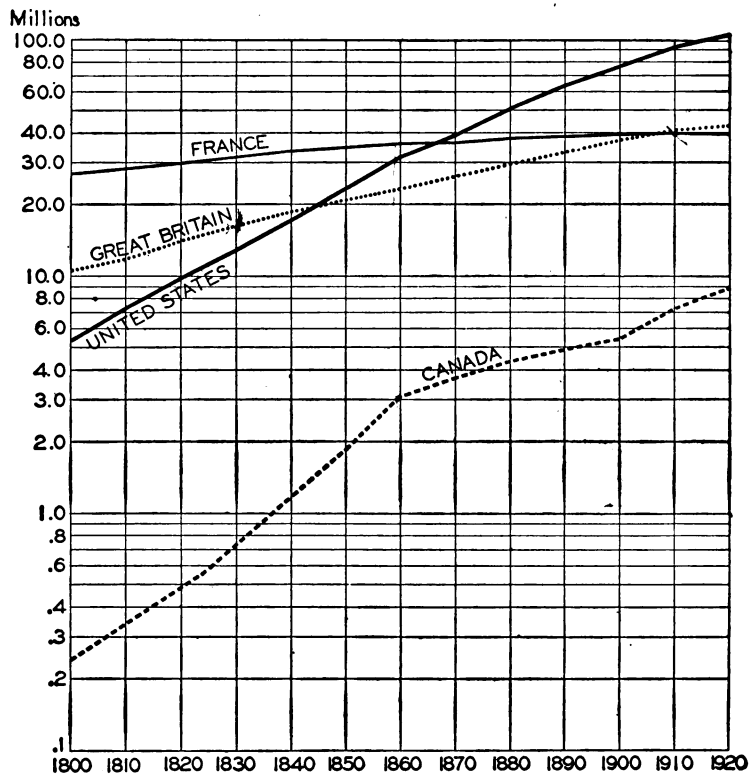


DIAGRAM VIII. POPULATION OF GREAT BRITAIN, FRANCE, CANADA, AND THE UNITED STATES, SHOWN ON THE "RATIO CHART"

(2) Rough measurement of the two slopes referred to in (1) shows that the rate of increase for the United States was about twice that for Great Britain.

(3) The downward curve of the graph for the United States after 1860 shows that the rate of increase gradually diminished after that date.

(4) This form of chart, unlike the relative number chart (Diagram VII), still permits some comparison of absolute values, even

Digitized by Google

though the attention and emphasis is on rates of change. It is of interest to see just when the United States passed Great Britain and France, for instance, while we are comparing rates of change. We also see that Canada is always much smaller than the other three, a fact which is completely lost sight of in Diagram VII.

(5) The points noted in (1), (2), and (3) could of course also be seen by direct analysis of the rates of change. The point is that this chart almost does the analysis by itself. Even if it later seems desirable to carry through more careful analysis, an inspection of the plot on the ratio chart will often be of value in showing where interesting relationships exist.

There are two types of problem in which it is especially desirable to use the ratio chart. The first is the determination of the general course of an historical series. If such a series does not seem to be approximately a straight line when graphed on ordinary scales, it is often worth while to graph it on the ratio chart, to see if it is perhaps subject to the law of constant percentage of change, or, as it is sometimes called, the compound interest law.

The second type of problem in which it is desirable to use the ratio chart is the comparison of historical series in whose relative rates of change we are more interested than in their absolute values. For instance, Mr. P. A. Bivins [1] was interested in comparing the relative fluctuations in the price of lead and tin. The use of relative numbers would in such cases, as in the above population chart (Diagram VII), unduly emphasize the year taken as base, by referring all other prices to that year. The ratio chart was found to give just what was wanted. Again, it often happens that a store would like to discover whether changes in amount spent for advertising precede corresponding changes in volume of sales. If the two quantities are plotted on the ordinary scale, changes in advertising are minimized because of the small relative size of advertising expenses as compared with sales. If the figures are plotted on the ratio chart, however, a 20 per cent change in advertising occupies

[1] Cf. article in *Industrial Management*, July, 1922, by P. A. Bivins, "Charting as an Aid in Stabilizing Profits."

Digitized by Google

as much space as a 20 per cent change in sales, and comparison is readily possible.

As the ratio chart does not permit a direct comparison of the *differences* between quantities, it should not be used when differences are to be stressed rather than ratios. Often the same information should be plotted both on ordinary and on ratio paper, to facilitate analysis from both points of view.

Principle of the ratio chart. If the practical statistician understands thoroughly the points discussed in the preceding section, he can make considerable use of the ratio chart forms published by various agencies, without going more fully into the underlying theory. He will use this method with more confidence, however, and probably also in a larger number of problems, if he understands why it is that this ruling does actually give the ratio property. The purpose of the following section, therefore, is to explain the fundamental principle which underlies the construction of these charts and leads to the ratio property. As this principle is merely that of logarithms, it will be familiar to many readers; but even for them it will perhaps be profitable to review the matter from the point of view of this particular application.

Suppose we think how successive points on the vertical scale of the ratio chart could be located. Let us say that the base of the chart corresponds to the number 1, and that a point a convenient distance above this corresponds to 10. The problem then is to determine the distances of other points above the base line, so that they will correspond to a series of other numbers, still keeping the fundamental ratio property. Let us try to get at the problem by locating with reference to 1 and 10 points which correspond to numbers such as 10^2, $\sqrt{10}$, etc. With these located we may be able to complete the scale so as to locate other numbers which are not simple powers and roots of 10.

Let us first, then, determine where the point representing 100 should be located. We note that the ratio of 100 to 10 is exactly the ratio of 10 to 1; therefore by the ratio property, the distance from the point representing 10 to the point representing 100 is the same as the distance from the point representing 1 to the point

Digitized by Google

representing 10. In other words, the point representing 100 is the same distance above the "10" as the "10" is above the "1," in this case the base line.

A similar argument shows that "1000" will be as far above "100" as "100" was above "10," and that $\frac{1}{10}$ will be that same interval below "1." We may note further that the distance from a number N to a number 10 N will, by this same ratio property, be the same as the distance from 1 to 10. If therefore we find the appropriate distances above the base line for all numbers N from 1 to 10, the distances above the base line for the numbers 10 N from 10 to 100 will be uniformly larger by the distance from 1 to 10. The scale from 10 to 100 looks exactly like the scale from 1 to 10, as far as the spacing is concerned; the difference is that the numbers represented are ten times as large. The ratio chart therefore consists of one, two, three or more banks of identical rulings.

The problem therefore reduces to finding the intervals between 1 and 10. A simple point to locate in that range is the square root of 10. By definition of square root, the ratio of 10 to the $\sqrt{10}$ is equal to the ratio of $\sqrt{10}$ to 1. The distance from the point representing $\sqrt{10}$ on the ratio chart to the point representing 10 is therefore the same as the distance from 1 to the point representing $\sqrt{10}$. In other words, the point representing the $\sqrt{10}$ will be halfway between "1" and "10." Similarly the point representing the square root of any number is halfway between the point representing that number and "1." We can apply this last statement to determining a series of correspondences, getting the points corresponding to the square root of $\sqrt{10}$ (i.e., $\sqrt[4]{10}$), then to the square root of this result, and so forth. Taking as our unit the actual distance of "10" above the base line, we have the following table of points on the scale and corresponding numbers.

It is possible to find a large number of other points on the scale by a combination of the points in this table. For instance, what is the number corresponding to the point .53125? This point is .03125 above the point .5, corresponding to $\sqrt{10}$, and the number corresponding to it bears the same relation to $\sqrt{10}$ as the number corresponding to .03125 bears to 1, by the fundamental property of the ratio chart. The desired number is therefore the product of the $\sqrt{10}$ and $\sqrt[32]{10}$, which turns out to be 3.398 2083.

Digitized by Google

Number Represented	Distance above Point Representing 1
10	1
$\sqrt{10} = 3.162\ 2777$	.5
$\sqrt[4]{10} = 1.778\ 2794$	.25
$\sqrt[8]{10} = 1.333\ 5214$	.125
$\sqrt[16]{10} = 1.154\ 7820$	.0625
$\sqrt[32]{10} = 1.074\ 6078$	.03125
$\sqrt[64]{10} = 1.036\ 6329$	.015625
.....	.....
1	0

It is obvious that even after we have found a large number of other points by this method, the results are not in convenient form for actually constructing a ratio chart. For that purpose the "Numbers represented" should be integers and numbers with only one or two decimal places. By a process known as "interpolation," it is possible to get the results in this form. Fortunately for us, the detailed computations were done a couple of centuries ago by Briggs, who used the term "logarithm to base 10" for what we have called "distance above point representing 1." One of his methods of computation, however, followed the lines here indicated.

A short table indicating the correspondence of "Numbers" and "Distances" is inserted here, so that the student may construct for himself a rough outline ratio chart.

Number	Distance	Number	Distance	Number	Distance
10	1	4.5	.6532	2	.3010
9	.9542	4	.6021	1.8	.2553
8	.9031	3.5	.5441	1.6	.2041
7	.8451	3	.4771	1.4	.1461
6	.7782	2.5	.3979	1.2	.0792
5	.6990	2.2	.3424	1	.0000

For some purposes it is convenient to locate the points according to their "distances" or logarithms, without definitely drawing up the grid of the ratio chart. This should be done when numerous horizontal lines would be confusing, in diagrams of a simple type intended to show relations in a broad way. In such cases the points can be located by finding from a table of logarithms the

Digitized by Google

logarithms of the numbers, and plotting these logarithms rather than the numbers. For rapid work, especially in the preliminary analysis of a problem, the use of the prepared grid is preferable.

Common errors in the use of percentages. (*a*) Ambiguity as to what is the base. — A trial question to ask after making or hearing a statement about percentage, is, Per cent of what? If the answer is at all uncertain, revise the statement. A merchant says for instance that his margin is forty per cent. He means 40 per cent of selling price, but many of his hearers think he means 40 per cent of the price he had to pay. In stating percentages such as these, do not be afraid of repetition, if desirable for clearness; in statistical English, grace of diction and even brevity are secondary virtues as compared with clearness, precision, and directness.

(*b*) Balancing a certain per cent increase against the same per cent decrease. Some persons seem to think that if wages are decreased 25 per cent say, and then increased 25 per cent, that they are back where they started, whereas each per cent is based on the wages at the time of change. The increase is therefore 25 per cent of 75 per cent of the original figure, and the net result 93.73 per cent of the original figure.

(*c*) Using percentages of increase of a small quantity as if comparable with percentage increase in a large quantity. — A new suburb may easily increase 100 per cent in a decade, whereas a larger unit can only do so under very unusual circumstances. Or the exports or imports of a certain commodity may increase by huge percentages when trade in it is just beginning. In such cases, the values as well as the ratios should be stated.

Summary of Chapter II. One of the first steps in statistical analysis, after statistical series have been constructed, is to find the *ratio* between two terms in the same or related series. This ratio is often expressed as a percentage to facilitate rapid comprehension.

The process of division required to reduce ratios to the percentage form is usually best performed on a computing machine, although, in certain cases, the slide rule, logarithms, tables, or long

Digitized by Google

hand division may be satisfactory, or under certain circumstances preferable.

Ratios (or percentages) may be used:

(1) To show the relation of parts to the whole;

(2) To measure progress relative to a standard;

(3) To show whether one quantity is large or small relatively to another quantity which has some logical or usual relationship to the first quantity;

(4) To express the terms of a historical series in relation to one such term ("relative numbers").

Comparison of ratios rather than absolute values is facilitated by the use of the *ratio chart*, in which the vertical scale is based on the logarithms of the quantities rather than the quantities themselves. The fundamental property of this chart is that the vertical distance between two points in it measures the ratio of the quantities represented, whereas in the ordinary graph it measures the difference between the two quantities. This chart is especially useful for plotting historical series in which each term exceeds the previous term by about the same percentage, and for comparing the relative changes of two or more historical series.

Certain common errors can be avoided by careful attention.

PROBLEMS

1. Study the attendance figures at your college for a series of years. Show by the use of ratio analysis the present and past composition of the group, as divided by classes, degrees sought, age, etc., and also show how these or other figures have changed from year to year.

2. Analyze the expenditures of the United States Government over a series of years.

3. Compare the price changes during the past month of United States Steel stock on the New York Stock Exchange and of wheat at Chicago.

Use the *Statistical Abstract of the United States* or the *World Almanac* as a source book for the following question.

4. Compute for each State in the Union the ratio (*a*) pupils per teacher in elementary schools; (*b*) teachers per 100 pupils in high schools; (*c*) pupils per teacher in high schools; (*d*) teachers per high school or academy.

5. Compute the proportion of the population of the United States in each State.

Digitized by Google

6. Compute the proportion of number of incomes, of amount of income, and of income tax paid, in each income group ($1000–$2000, $2000–$3000. etc.), for the latest year for which figures are available.

7. Compute the series of relative numbers for the population of Rhode Island, Washington, and the United States, for each decade from 1860 to 1920, first with 1860 as base year, and then with 1920 as base year. Graph each of these sets of series.

8. Graph the populations of Rhode Island, Washington, and the United States from 1860 to 1920 on the ratio chart, and discuss the relative advantages of this chart and that found in the preceding problem.

9. A speaker said that the general price level rose 100 per cent then declined 20 per cent and was then 60 per cent above its starting point. Explain his statement.

Digitized by Google

CHAPTER III

THE COMBINATION OF RATIOS

In the previous chapter, we have discussed various ways in which ratios (usually expressed as percentages) may be used to bring out more clearly the significance of statistical series and to facilitate comparisons. Important problems arise when we try to connect the ratio (e.g., death rate) which applies to a large group with the corresponding ratios which apply to the sub-groups of which the group is composed.

For instance, consider a simple question along this line — how may we infer the percentage which applies to the entire group when we know the percentages which apply to the sub-groups? Or more briefly, how may we combine percentages?

A correct method for combining percentages. As a typical example, let us suppose that on the first of July a baseball team has won 42 games out of a total of 67; its "percentage" is therefore $\frac{42}{67} = .627$. The next week it wins 3 out of its 7 games, and the percentage applying to that week's games is therefore $\frac{3}{7} = .429$. How can we find the percentage applying to all games played at the end of the week, by means of some combination of the two percentages stated?

We can find the final percentage correctly as follows: Add the 3 new wins to the original 42, and the 7 new games played to the original 67; the totals are 45 and 74, and the percentage won is $\frac{45}{74} = .608$. If we choose, we may write this computation:

$$\frac{67 \times \frac{42}{67} + 7 \times \frac{3}{7}}{67 + 7} = \frac{42 + 3}{74} = \frac{45}{74} = .608$$

We may write further, instead of the first fraction:

$$\frac{67}{74} \times \frac{42}{67} + \frac{7}{74} \times \frac{3}{7} \qquad \text{or} \quad \frac{67}{74}\,.627 + \frac{7}{74}\,.429$$

As all of the above expressions are merely numerical equivalents,

Digitized by Google

it is clear that we may obtain the combined ratio by multiplying each ratio by the proportion of cases on which it is based and adding all such products, that is to say, by multiplying .627 by $\frac{4}{7}$ and .429 by $\frac{3}{7}$, and adding the two products.

It is also correct, though usually less convenient, to multiply each ratio by the *number* of cases on which it is based, add the products, and divide by the total number of cases.

In either method, the numbers which are used to multiply into the ratios are known as *weights*. Using this term, we can restate the process as follows:

To combine the ratios applying to sub-groups, multiply each ratio by the appropriate weight and add all such products. The appropriate weight for each ratio is the proportion of all cases which is included in the corresponding sub-group.

Algebraic proof of correctness of above method. An algebraic proof of the correctness of the above rule is given below, partly for the purpose of introducing the use of algebraic symbols to aid the discussion.

Let $t_1, t_2, \ldots t_n$ denote the total number of cases in the first, second, etc., of n sub-groups;

and let $a_1, a_2, \ldots a_n$ denote the number of favorable cases, i.e., the number of cases in which a certain specified result occurs, in the first, second, etc., sub-group.

Then the ratio of the number of favorable cases to the total number of cases is

$a_1/t_1, a_2/t_2, \ldots a_n/t_n$, respectively, for the sub-groups,

and $\dfrac{a_1 + a_2 + \ldots + a_n}{t_1 + t_2 + \ldots t_n}$ for the total of all sub-groups.

The "appropriate weights," moreover, are obviously

$$\frac{t_1}{t_1 + t_2 + \ldots + t_n} \quad \frac{t_2}{t_1 + t_2 + \ldots + t_n} \quad \ldots\ldots\ldots \quad \frac{t_n}{t_1 + t_2 + \ldots + t_n}$$

respectively.

Digitized by Google

Now according to the rule, the "combined" ratio of favorable to total cases is

$$\frac{t_1}{t_1+\ldots+t_n}\ \frac{a_1}{t_1}+\frac{t_2}{t_1+\ldots t_n}\ \frac{a_2}{t_2}+\ldots\ldots+\frac{t_n}{t_1+\ldots+t_n}\ \frac{a_n}{t_n}$$

This expression can be simplified at once to exactly the form stated above as the ratio of favorable to total number of cases for the total of all sub-groups, thus justifying the rule.

It should be noted that the weights can be determined just as well if the *proportion* of cases in each sub-group is known, even if actual numbers are not known.

"Averaging in" a percentage based on additional cases. Using the same argument, we can justify the following short cut for averaging in additional cases. Reverting to the numerical example, we may express the resulting percentage as

$\frac{7}{74}.627+\frac{7}{74}.429=.627-\frac{7}{74}(.627-.429)=.627-.0946\,(.198)$
$=.627-.019=.608.$

Or in symbols: To average together two ratios, r_1 and r_2, based on t_1 and t_2 cases, respectively, add to r_1 the quantity $\frac{t_2}{t_1+t_2}(r_2-r_1)$.

This gives $r_1+\frac{t_2}{t_1+t_2}(r_2-r_1)=\frac{r_1t_1+r_1t_2+t_2r_2-t_2r_1}{t_1+t_2}=\frac{r_1t_1+r_2t_2}{t_1+t_2}$,

which agrees with the fundamental rule for combining ratios.

Or in words, to average together two ratios, find the quotient of the number of cases in the second group, divided by the total number of cases; multiply this quotient by the difference between the second ratio and the first ratio; if the second ratio is greater than the first, add the result to the first ratio; if the second ratio is less than the first, subtract the result from the first ratio.

Example: A team has played 50 games for a standing of .580; they lose the next game; what is their standing? Here t_1 is 50, t_2 is 1, r_1 is .580, $r_2=0$. The standing is therefore $.580+\frac{1}{51}(0-.580)=.569$.

Digitized by Google

Effect of changing weights in combining ratios. It is obvious that if in two different instances it is necessary to combine the same ratios, the resulting ratios will ordinarily be different. Reverting to our original example, let us suppose a second instance, in which a team has a percentage of .627 for 67 games, and .429 for 14 games. The combined percentage is then $.627 - \frac{14}{81} \times .198 = .593$, as compared with .608 in the original instance. In such instances as these, the composite percentage will vary, according to the relative distribution of the cases on which the component ratios are based, from something nearly as large as the larger ratio to something nearly as small as the smaller. For instance, if a ratio of .600 is based on 1000 cases, and one of .400 on 10 cases, the combined ratio is .598; if the .600 ratio is based on 10 cases, and the .400 ratio on 1000, the combined ratio is .402. If three or more rather than two ratios are combined, the composite ratio may vary from nearly as large as the largest to nearly as small as the smallest, according to the distribution of cases.

Algebraic discussion of composite percentages. In this numerical illustration, it is clear that the weights are of major importance in determining the composite percentage which applies to a large group from the individual percentages which apply to its component sub-groups. As this process of determining such a composite percentage is of the greatest importance in statistical work, it is desirable to formulate in general terms some principle which will help us to decide when differences in weights will make a real difference in the resulting composite percentage, and when such differences are probably not significant.

In order to formulate such a principle in a general way, it is desirable to state the formula for a composite percentage in algebraic language, and to reduce this formula to its simplest form. It was stated *in words* in the first part of this chapter, that "to combine the ratios applying to sub-groups, multiply each ratio by the appropriate weight, and add all such products. The appropriate weight for each ratio is the proportion of all cases which is included in the corresponding sub-group." The algebraic notation used above may be abbreviated by letting:

Digitized by Google

$p_1, p_2, \ldots, p_n = \frac{a_1}{t_1}, \frac{a_2}{t_2} \ldots \frac{a_n}{t_n}$, the component ratios;

and $m_1, m_2, \ldots, m_n = \frac{t_1}{\Sigma t}, \frac{t_2}{\Sigma t} \ldots \frac{t_n}{\Sigma_t}$, the proportion of all cases in the first, second, ... and n'th sub-group.[1] It is obvious, then, that $m_1 + m_2 + \ldots + m_n = \frac{t_1 + t_2 + \ldots + t_n}{\Sigma t} = 1$, and that, in accordance with the principle just quoted,

$$P \text{ (the composite ratio)} = m_1p_1 + m_2p_2 + \ldots + m_np_n.$$

Let us apply this formula in a case in which the m's and p's are given, although the a's and t's are not stated. Let us suppose the death rate for colored people in a certain country is given as 40 per thousand per year, and the death rate for whites as 20 per thousand per year, and nine tenths of the population are known to be whites. We have then $p_1 = .04$, $p_2 = .02$, and $m_1 = .1$ and $m_2 = .9$, and the computation of P is:

$$P = .1 \times .04 + .9 \times .02 = .022, \text{ or } 22 \text{ per thousand.}$$

As the recorded facts on which computations must be based are often given as ratios rather than in full detail, this method of calculation is often serviceable.

Algebraic discussion of effect of differences in weights. If a second composite percentage is formed in which the component ratios are the same as in the first case, while the weights are different, we may state algebraically the magnitude of the difference, as follows: Let $m_1', m_2', \ldots, m_n'$ be the second set of weights (proportions of cases), and P' the resulting composite ratio. Then as above we have:

$$P' = m_1'p_1 + m_2'p_2 + \ldots + m_n'p_n$$

Then the difference between the two composite ratios is given by:

$$P' - P = (m_1' - m_1)\, p_1 + (m_2' - m_2)\, p_2 + \ldots + (m_n' - m_n)p_n$$

As we know that $\Sigma m = 1$ and $\Sigma m' = 1$, we have:

[1] The Greek letter Σ placed in front of a letter or expression means "the sum of all the." In this case, Σt means "the sum of all the t's."

Digitized by Google

$$\Sigma m' - \Sigma m = 0,$$

or rearranging terms,

$$(m_1' - m_1) + (m_2' - m_2) + \ldots + (m_n' - m_n) = 0$$

In words, this last equation says that if some of the differences of corresponding weights are positive, other differences must be negative. The value of the difference $P' - P$ therefore depends on which p's are multiplied by the positive differences of the m's and which by the negative differences.

If the positive values of these differences are on the whole multiplied into large values of the p's while the negative values of the differences are multiplied into small values of the p's, the positive terms on the right will on the whole be larger than the negative terms, and $P' - P$ will be plus, i.e., P' will be larger than P. Conversely, if positive values of $m' - m$ are on the whole associated with the smaller p's, and negative values of this difference with the larger p's, the negative terms on the right will be larger than the positive terms, and P' will be less than P. If on the whole there is no tendency for the larger p's to be multiplied into either positive or negative values of these differences, and especially if the number of p's is considerable, $P' - P$ will usually be relatively small as compared with P.

This principle in this form seems perhaps obvious and almost trifling, but the applications are, as a matter of fact, of major importance. The reason for this importance lies in the fact that many ratios commonly used are really combinations of simpler ratios. The "percentage" of a baseball team is a combination of its "percentages" against seven different teams. The operating ratio of a railroad is a combination of such ratios for its different divisions for different months. So for birth rate, death rate, marriage rate, rate of expenditure of ammunition by guns in battle, or in fact any ratio of any generality. The problems raised by such combinations of ratios cannot always be solved by going back to the constituent ratios. It is necessary to have a sound understanding of the way in which changes in the factors will affect the results.

A very good rate to study to clear up this matter is the death rate, as the death rate of a city is a combination of rates for various

Digitized by Google

age-groups, races, two sexes, and so forth, and the factors are the proportions in each of these groups and vary considerably from city to city. It is easy to see, as an application of the preceding discussion, that two cities might have exactly the same "specific" death rates — e.g., for white males aged 30–35, and every other group — but because of the different composition of the population in the two cities: i.e., the different m's, have considerably different general death rates.[1]

Because of the fallacy illustrated by this example, Whipple[2] makes the statement that "the way to use a general death rate is to analyze it." A way to meet the problem of presenting the facts as to death rate in a single figure, but yet avoid this fallacy, is to use instead of the crude death rate, the death rate of a standard population — i. e., one in which the m's are taken to be, not what they are for the particular political unit considered, but certain typical values that have been agreed on, for instance, the proportions of population in the country as a whole which fall in each age group. This same general idea can be applied in other fields, subject always to the limitation that it will often not be understood without considerable explanation. In the opinion of the author, the dangers of misapplication should not lead to complete disuse of general rates; the point is rather that they should be used only with due precautions, and a realization that *two such rates are comparable only if the proportions of the constituent parts are substantially the same.*

Other instances of the application of the principle that a combined ratio is changed by an alteration of the m's, are indicated in the following examples.

EXAMPLES. **1.** Show by a numerical example that a baseball team A might have a better standing than a team B against every other team in the league, and still have a lower standing as a whole.

2. Show that a railroad with five divisions might have a lower cost per ton-mile in July than in June on every division, and yet have a higher cost per ton-mile for the railroad as a whole.

---

[1] See, for example, Whipple's *Vital Statistics*, 2d edition, p. 285, where it is shown that Cambridge, Massachusetts, with actually lower specific death rates, had a higher general death rate than Chicago, Illinois.

[2] Reprinte l by permiss on from *Vital Statistics*, by G. C. Whipple, and published by John Wiley and Sons, Inc. Second Edition, 1923, p. 351.

Digitized by Google

3. A Japanese speaker argued recently that the apparently high birth rate of the Japanese in California was due to the fact that an unusually large proportion of the Japanese population was between the ages of 15 and 45, and that later this high birth rate would be reduced as the age distribution of the population became more normal. Discuss the validity of this argument.

4. Two salesmen, Jones and Smith, sell three types of goods, A, B, and C, on which the percentage of profit to the firm is 30, 20, and 10 per cent, respectively. Make up a numerical schedule showing Jones selling goods with a greater total selling price, but smaller total profit to the firm than Smith, and show that this is an instance of shifting weights in combining ratios.

"Homogeneous" material. The preceding discussion has a general bearing on the work of a statistician in any line. We have seen that the general death rate may produce an absolutely false impression. It is agreed in such cases that a trustworthy statistician should not issue figures of that sort. That is to say, before publishing any totals, percentages, averages, etc., he is supposed to have examined every subclassification of the material and to have made sure that the proportions of the different constituents are not different enough from what would normally be found, to invalidate comparisons that would naturally be made. The error guarded against has been styled that of "concealed classification," [1] and statistical purists have sometimes said that only "homogeneous" material should be used as a basis for totals, averages, percentages, etc. They have neglected, however, to define what they mean by homogeneous. They might mean that the material must be so analyzed that all units or individuals assigned to the same class or group are exactly alike. This definition of homogeneous would be satisfied in the exact sciences; there are no differences between two different gallons of water, for instance. The essential point of the social sciences, however, is that different individuals are not alike, and that while they are being classified on the basis of two or three qualities or measured characteristics, the existence of differences in other characteristics is temporarily minimized but not ignored. It is therefore an essential part of statistical analysis that the material must *not* be "homogeneous" in the sense suggested above.

Let us consider an instance of simple statistical classification — the standing of one of the baseball leagues. The units classified are

[1] Whipple, p. 238.

Digitized by Google

baseball games, and these units are divided into groups on the basis of which team won the game. For the purpose in hand, this classification is unquestionably valuable, but it can certainly not be said that all the games thrown into one group because won by team A are exactly alike. In the first place, they were played against different opposing teams; they were played in different cities, in different kinds of weather, at different times in the year, and before different groups of spectators. The real point of the classification is that these differences are not significant in determining the "standing" of the different teams, according to the recognized rule. Furthermore, on the whole, these differences do not favor one team more than another, so that the "standing" may be expected to correspond to relative excellence. The baseball enthusiast will examine these differences and perhaps discover, especially during the earlier part of the season, that the "standing" is not a fair measure of the relative worth of one of the teams. That team has perhaps had more games with the best of the other teams than a second team which stands higher in the general percentage of games won. The criticism which such an enthusiast would make of the published percentages is in our opinion the same in principle as the criticism of those who object that certain material is not homogeneous. The point is, not that the units summarized in one class differ from each other in various characteristics, but that these differences are unusual enough to invalidate the conclusions which would ordinarily be drawn.

We believe, then, that in a strict sense statistics cannot be expected to be homogeneous, but that the idea in the minds of those who use this phrase is fundamental to sound statistical work. It is incumbent on the statistician to draw up his classes in such a way that the individuals in a group shall be, as far as possible, similar to each other, or at least, that the differences in other characteristics than the ones made the basis of classification, shall not be exceptional. If they are exceptional, any conclusions which may be reached will apply only in other cases in which these same peculiarities exist.

From this point of view, it is clear that the task of the statistician is very exacting. The bases for classification which he uses must

Digitized by Google

be the most important for the purpose in hand, and must include all classifications which might reasonably be expected to affect his conclusions. Much detailed work must be done, in running down possibly important factors, which will often not appear explicitly in the results. On account of the complex interrelationships of social problems, the analysis of any one of them calls for broad general knowledge, and an unprejudiced and keen mind. Any one who aspires to be a statistician, and not merely a juggler of figures, must test his groups for possibly important differences which have been overlooked, and weigh all his tentative conclusions in the light of such tests.

In particular, any general ratio, such as the death rate, or the operating ratio, should be examined in the light of available sub-ratios to determine the conditions under which it is valid. If we realize, for instance, that a certain death rate was found for a population which included 20 per cent of negroes and an unusually large proportion of whites over 60, we can avoid many unfair comparisons. Again, a realization that differences in over-all ratios may be due to different relative importance of different sub-groups rather than to real differences in the sub-ratios for the groups, will help to a better interpretation of such differences when found. A difference between two values of a general ratio should therefore not be taken as conclusive evidence of significant differences in component ratios, but merely as a signpost indicating that there are either such differences or differences in the relative weights. Any statistician who submits such a difference as significant is supposed to have examined the various possibilities that the apparent result is due to a shifting of weights in the sub-groups, or in other words, to have made sure that his material is "homogeneous."

Summary of chapter. The value of a ratio for a group made up of several sub-groups may be obtained by multiplying each sub-ratio by the number of cases on which it is based, adding these products, and dividing the total by the total number of cases.

Modified or abbreviated forms of this process facilitate the combination of ratios.

When two general ratios are obtained from the same sub-ratios

Digitized by Google

by combining them by means of two different sets of weights, the two general ratios will be substantially different if the larger sub-ratios are on the whole multiplied by more important weights in one case than in the other.

On account of the great frequency with which supposedly comparable general ratios are composed of sub-ratios multiplied by significantly different weights, all general ratios should be used with great caution. Two such ratios are comparable if and only if the proportions of the constituent parts are substantially the same.

A full understanding of the preceding point really covers the real meaning of the general principle enunciated by some statisticians that all statistical analysis should be based on homogeneous material. Any statistician who presents results should have examined his data for unusual distribution into classes, and should call attention to any peculiarities discovered. More especially, he must be sure that any supposedly significant differences in ratios are real and not due to a shifting of weights among the sub-ratios.

Digitized by Google

CHAPTER IV

FREQUENCY AND QUANTITY DISTRIBUTIONS

We have seen that the first step in statistical analysis is the formation of statistical series, by grouping items on the basis of one or more principles of classification. The formation of such series implies of course the statement of results in the form of tables, and by such graphic devices as involve merely picturing the figures found. A second step in statistical analysis is one or another of the forms of ratio analysis which have just been discussed, in the two preceding chapters, supplemented by appropriate graphic presentation. The further steps of analysis are not, like these two, much the same for all types of series, but involve specialized methods. Most economic and business statisticians are more familiar with the methods suitable for the analysis of historical series, while biometric and educational statisticians use methods founded on frequency distributions. This division of the field will not, in the opinion of the author, be true permanently except for the lower grades of statisticians. The thorough understanding of frequency distributions is of fundamental importance for a statistical understanding of the universe. Statistics has been defined as "the science of averages"; a better definition would be "the science of frequency distributions." No business statistician, then, can afford to slight the methods used with frequency distributions.

Many questions in ordinary life which are actually answered in a rough way as "so and so on the average" or "about so much" are more satisfactorily answered by a frequency distribution. How old, for instance, are freshmen when they enter college? We may answer, 18 years and 10 months on the average, but it is certainly more informing to say that the number at each age for a typical class of 400 was as follows:

Age	15	16	17	18	19	20	21	22
Number	1	6	49	130	90	31	11	2

Digitized by Google

Such a complete series enables one to see how closely the bulk of the students approximate the average, how many extremely divergent cases there are, and how many are younger than any given age.

Other questions which are better answered by a complete frequency distribution than by an average are the following:

Age of graduating class.
Wages of laborers, or any group of skilled workers.
Size of class-section, in college or school.
Number of passengers carried per trip, on a street car.
Time of arrival of a train at a station, on various days.
Minutes a person takes to eat breakfast — lunch — dinner, on various days.
Time required to prepare a lesson.
Miles covered per day, by an ocean liner.
Age at marriage, of any group of men or women.
Miles covered per gallon of gasoline, by an automobile.

For many purposes, of course, the rough "average" is good enough; but when questions like these arise in the course of a business problem or a scientific investigation, it is often worth while to make considerable effort to get the full frequency distribution. And, in many other cases, it is imperative to realize that there is a frequency distribution back of the summary figure used.

A statistician should be able to grasp the significance of frequency distributions promptly and also to present those which arise in his work in such form that his readers or hearers can easily comprehend them. As one means of doing this he should use one or more of the graphical schemes which have proved successful in similar problems. In order to bring out this point, we shall show by each of the more important methods the treatment of the frequency distribution of the weight of men 5 feet 8 inches tall between ages 20–24, as found in the *Medico-Actuarial Mortality Investigation*.[1] In several of the methods, we shall show also the graphs of several other distributions, to bring out particular points, and in some cases, we shall show for purposes of comparison in the same diagram the curves for the distribution of weight in men of different heights or ages. The graphic representation of frequency distribution should

[1] Compiled by the Association of Life Insurance Medical Directors and the Actuarial Society of America, and published under the title, *Medico-Actuarial Mortality Investigation*, New York, 1912–15.

Digitized by Google

not be thought of as something interesting but nonessential; it is in fact a part of the mechanism of thought about them, without which it would be difficult to have clear ideas and also awkward to express them.

The histogram. Let us take up, then, the sample frequency distribution mentioned, that is to say, the weights of those men included in the investigation who were 5 feet 8 inches in height and 20–24 years of age. It was found that the number of men in each specified weight group was as follows:

Weight	108–122	123–137	138–152	153–167	168–182	183–197	198–212	213–227
Number of men	74	1410	2712	1148	334	68	17	3

It may be noted in passing that this frequency distribution is typical of most that occur; that is to say, that the number of items in an interval is small for low values of the quantity, is greater for larger values, and for still larger values again becomes small. This particular distribution is less symmetrical than many.

The first form of graphic representation for such data is the *histogram;* we construct a series of vertical columns of equal width

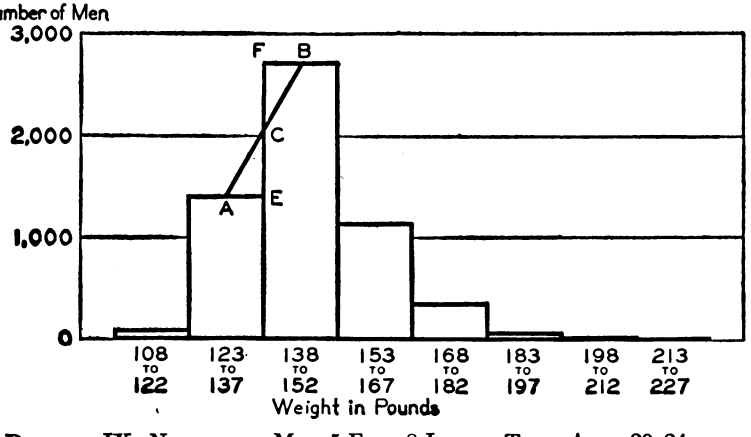


DIAGRAM IX. NUMBER OF MEN 5 FEET 8 INCHES TALL, AGED 20–24, IN EACH INDICATED WEIGHT GROUP

Digitized by Google

and of height proportional to the number in each group; indicate on the vertical scale what these numbers are, and on the base, at the boundaries of the columns or at the mid-points, indicate the weights.

If the class intervals are not uniform in width, it is necessary to modify this method somewhat, the guiding principle being that a unit of area should always represent the same number of items. If therefore one class interval is five times as wide as another, the frequency for that interval should be divided by 5 before it is plotted. The ordinates are then proportionate to frequency-density rather than to frequency, although in the more usual case of equal intervals the two ideas are equivalent. This scheme of presentation is not suitable for use in diagrams which are to be shown or published to an audience without technical training, but is essential for clear ideas in discussion of the theory back of frequency distributions.

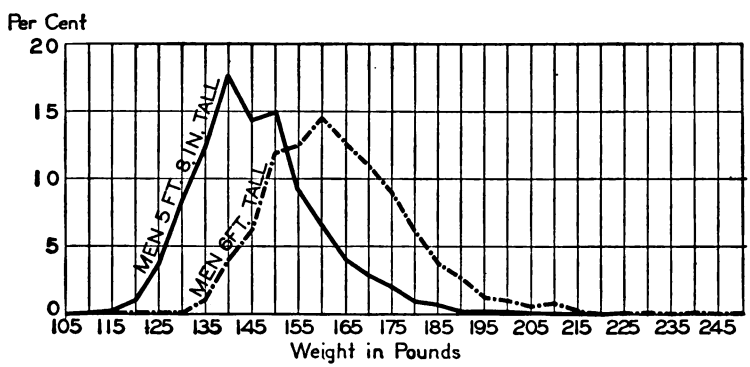


DIAGRAM X. PER CENT OF MEN 5 FEET 8 INCHES TALL, AND OF MEN 6 FEET TALL, IN EACH INDICATED WEIGHT GROUP
All figures are limited to men aged 20–24

The frequency polygon. A second form of diagram is somewhat more common. If the successive mid-points of the upper boundary of each bar in the histogram are joined by straight lines such as AB, there results a frequency polygon, whose area is just the same as that of the histogram, since triangle AEC is obviously equal to

Digitized by Google

PER CENT OF MEN 5 FEET 8 INCHES TALL, AND OF MEN 6 FEET TALL, IN EACH INDICATED WEIGHT GROUP

WEIGHT	MEN 5 FEET 8 INCHES TALL	MEN 6 FEET TALL
108–112	.07	.05
113–117	.21	
118–122	1.00	.09
123–127	3.72	.09
128–132	8.33	.09
133–137	12.36	1.07
138–142	17.76	3.95
143–147	14.42	6.23
148–152	14.94	11.95
153–157	9.30	12.47
158–162	6.53	14.51
163–167	4.05	12.65
168–172	2.87	10.98
173–177	2.03	8.89
178–182	.88	6.09
183–187	.74	3.82
188–192	.16	2.74
193–197	.28	1.26
198–202	.17	1.02
203–207	.05	.56
208–212	.07	.79
213–217	.04	.28
218–222		.19
223–227	.02	.09
228–232		
233–237		
238–242		.09
243–247		
248–252		.05

triangle *CBF*. In order to have this equality true at the upper and lower end of the figure, it is necessary to join the mid-point of the top of the last column with the mid-point of the next interval along the base. This polygon may therefore be drawn directly by plotting points at a height proportional to the frequency of quantities in

Digitized by Google

that interval, directly above the mid-point of the interval on the base line.

The method of doing this is illustrated by a diagram (Diagram X) drawn on this basis for essentially the same figures as shown on the histogram. We desire, however, to compare with the distribution previously shown a corresponding distribution with respect to weight of men of the same age but 6 feet tall rather than 5 feet 8. As the number of such individuals is 2150 as compared with 5776, no valuable comparison can readily be made by the rise of the actual figures. A better basis is the per cent of total in each class, which is therefore used instead of the actual figures. In order, moreover, to show the facts in more detail we use a 5-pound weight interval rather than the 15-pound interval used above.

As another example of a frequency distribution, let us take the distribution by age periods of the males and females in the United States in 1920, according to the Census returns. (Diagram XI.)

The shape of the frequency polygon summarizes the nature of the

DISTRIBUTION BY AGE PERIODS IN 1920 OF MALES AND FEMALES OF KNOWN AGES

AGE PERIOD	MALES		FEMALES	
	Number	Per cent	Number	Per cent
0 to 4	5,857,000	10.88	5,716,000	11.04
5 to 9	5,753,000	10.69	5,645,000	10.91
10 to 14	5,369,000	9.98	5,272,000	10.19
15 to 19	4,674,000	8.69	4,757,000	9.19
20 to 24	4,527,000	8.41	4,750,000	9.18
25 to 29	4,538,000	8.43	4,548,000	8.79
30 to 34	4,131,000	7.68	3,940,000	7.61
35 to 39	4,074,000	7.57	3,701,000	7.15
40 to 44	3,286,000	6.11	3,060,000	5.91
45 to 49	3,118,000	5.80	2,646,000	5.11
50 to 54	2,536,000	4.71	2,199,000	4.25
55 to 59	1,880,000	3.49	1,669,000	3.23
60 to 64	1,582,000	2.94	1,401,000	2.71
65 to 69	1,080,000	2.01	989,000	1.91
70 to 74	796,000	1.31	689,000	1.33
75 to 79	420,000	0.78	436,000	0.84
80 to 84	186,000	0.35	217,000	0.42
85 to 89	69,000	0.13	87,000	0.17
90 and over	22,000	0.04	32,000	0.06
Total	53,808,000	100.0	51,754,000	100.0

Digitized by Google

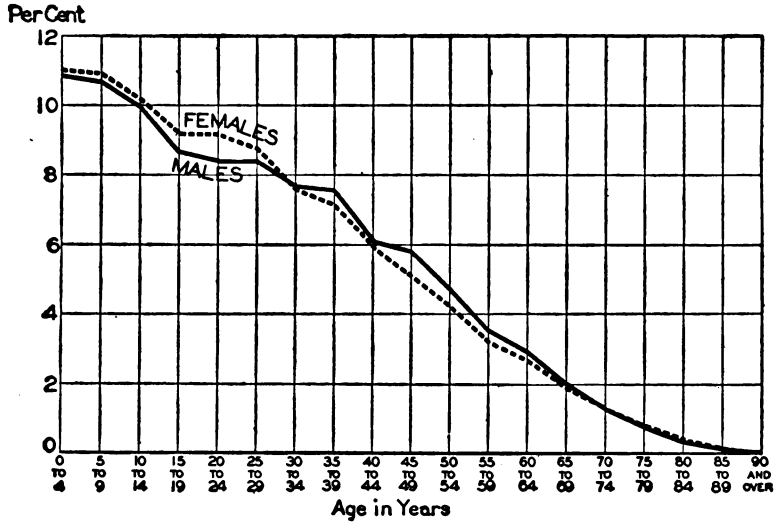


DIAGRAM XI. PERCENTAGE DISTRIBUTION BY AGE PERIODS OF MALES AND FEMALES IN THE UNITED STATES IN 1920

frequency distribution. An experienced statistician becomes familiar with certain typical forms of the polygons and becomes able to estimate approximately the characteristics of a distribution by a glance at its polygon. Among the most frequent forms are the following:

A. Polygons similar to those of Diagram XI, which run down from a high point on the left (or right) to zero on the right (or left). Such polygons have been described as "J-shaped." [1]

B. Polygons which are approximately symmetrical about the maximum ordinate. That is to say, they start with short ordinates to the left, rise to a peak, and then fall away again to the right, the heights at equal distances to the right or left of the maximum being the same.

C. Polygons similar to the preceding, except for the lack of symmetry. That is to say, the polygon falls away to the right and left of the maximum, but the height at any distance to the right is different from the height at the same distance to the left. Or in

[1] From *Introduction to the Theory of Statistics*, by Yule, pp. 98–102.

Digitized by Google

more informal terms, as we move from left to right, the polygon rises more rapidly to its peak than it falls away on the other side, or vice versa. This departure from symmetry is known as skewness, and occurs very frequently in the frequency distributions commonly encountered.

By increasing the number of class intervals in a frequency distribution, we can increase the number of vertices in the frequency polygon. As will be shown in later chapters, we may think of extending this process in such a way that the polygon approaches a certain curve as a limit, under conditions which will be specified in the later discussion. Among the theoretical curves which may be derived in this way are the three shown in Diagram XII. The bell-shaped curve called the "normal curve" corresponds to polygons of type B above, and the other two curves correspond to polygons of type C of different degrees of skewness.

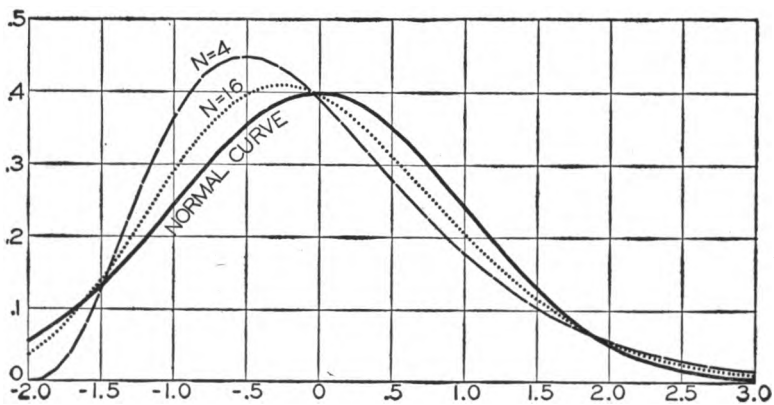


DIAGRAM XII. THREE THEORETICAL FREQUENCY CURVES

It should not be inferred that all frequency curves that are symmetrical are "normal," nor that all frequency polygons derived from "normal" frequency distributions will be symmetrical. The facts are merely that under certain conditions the limit of a series of polygons of type B (or in fact of some particular series of type C) will be exactly the "normal" curve as shown.

Digitized by Google

Cumulative diagrams. Another way of comparing two or more frequency distributions is to form what are known as the cumulative figures for each. Starting at the smaller end we add to the number (or per cent) in the first group the number (or per cent) in

Col. I Wt. in lbs.	Per Cent of Men Lighter than Weight Stated in Col. I	
	Men 5 ft. 8 in. tall	Men 6 ft. tall
112.5	.07	.05
117.5	.28	.05
122.5	1.28	.14
127.5	5.00	.23
132.5	13.33	.32
137.5	25.69	1.39
142.5	43.45	5.34
147.5	57.87	11.57
152.5	72.81	23.52
157.5	82.11	35.99
162.5	88.64	50.50
167.5	92.69	63.15
172.5	95.56	74.13
177.5	97.59	83.02
182.5	98.47	89.11
187.5	99.21	92.93
192.5	99.37	95.67
197.5	99.65	96.93
202.5	99.82	97.95
207.5	99.87	98.51
212.5	99.94	99.30
217.5	99.98	99.58
222.5	99.98	99.86
227.5	100.00	99.86
232.5		99.86
237.5		99.95
242.5		99.95
247.5		100.00
252.5		

Digitized by Google

the second. This sum is the number (or per cent) for which the value of the quantity is less than at the upper boundary of the second interval. We then add the number, in the third group, and note that the result is the number for which the value of the quantity is less than at the upper boundary of the third interval. And so we continue to the end of the table. We may derive from our previous figures, then, the percentage of men of each height whose weight was less than the figure indicated, as above.

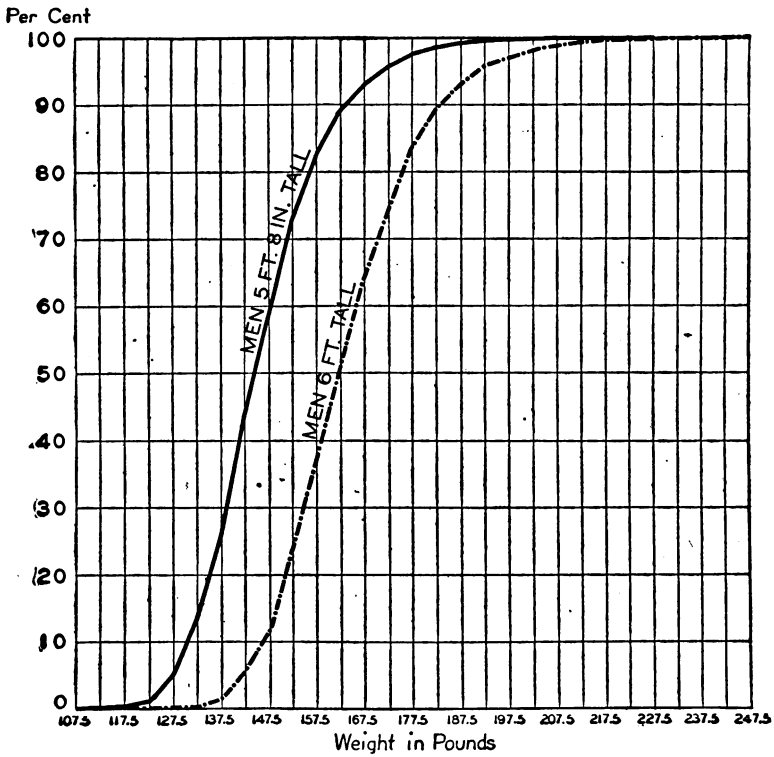


DIAGRAM XIII. PER CENT OF MEN LIGHTER THAN EACH WEIGHT INDICATED ON BASE LINE

These figures may be graphed as in Diagram XIII. It will be noted that the curves do not cross in the way that the frequency

Digitized by Google

polygons do, and permit comparison more clearly on the basis of two similar points. For instance, the fact that the percentage of men in the 160 lb. group is smaller for men 5 feet 8 inches tall than for men 6 feet tall, as shown on the polygons, is not significant except in the light of other facts. But the fact that 82 per cent of men 5 feet 8 inches and only 36 per cent of men 6 feet tall weigh less than 157.5, is a summary statement in itself of the comparative distribution.

It is of course equally desirable to cumulate the figures from the other end, getting successive figures for percentage of men whose weight exceeds such and such a figure. In either case, special care should be taken in drawing up the title and explanation of the diagram or table, as it is a less familiar type than the frequency polygon. It is superior, however, for comparing three or more frequency distributions.

"More than" diagrams. Special attention should be given to the form of cumulative distribution and its accompanying diagram obtained by summing from the upper rather than the lower end of the scale. The figures then express the per cent of all items for which the value of the quantity is *more* than each indicated figure. These diagrams are therefore known as "more than" diagrams although in particular cases it is clearer to say "heavier than," "wealthier than," or to use some other specific adjective. The advantage of this diagram as compared with the "less than" diagram is that the curve for a distribution in which the measured quantity on the whole runs large will lie *above* the curve for a distribution in which the measured quantity on the whole runs small, and the distance between the curves in a way indicates the general extent of the difference. A higher position for a curve representing larger quantities seems to most readers more natural than the reverse situation; and on the whole the "more than" diagram is perhaps more easily understood than the "less than" diagram.

It should be stated, in order to facilitate understanding of other texts and statistical articles, that "more than" and "less than" curves are often referred to as "ogives." Some statisticians prefer to use the horizontal axis for the percentages and the vertical axis for the variable. It is difficult, if not impossible if that is done,

Digitized by Google

however, to find a title for the diagram which is both simple and accurate, without violating the principle that the first noun in a title should give, if practicable, the meaning of the ordinates of the curves.

Col. I. Wt. in lbs.	Per Cent of Men Heavier than Weight Stated in Col. I.		
	Men aged 20–24, 5 ft. 8 in. tall	Men aged 20–24, 6 ft. tall	Men aged 35–39, 5 ft. 8 in. tall
107.5	100.00	100.00	100.00
112.5	99.93	99.95	99.97
117.5	99.72	99.95	99.90
122.5	98.72	99.86	99.36
127.5	95.00	99.77	97.82
132.5	86.67	99.68	93.90
137.5	74.31	98.61	85.72
142.5	56.55	94.66	75.64
147.5	42.13	88.43	65.13
152.5	27.19	76.48	53.22
157.5	17.89	64.01	43.99
162.5	11.36	49.50	34.22
167.5	7.31	36.85	26.09
172.5	4.44	25.87	19.51
177.5	2.41	16.98	14.05
182.5	1.53	10.89	9.62
187.5	.79	7.07	6.94
192.5	.63	4.33	4.79
197.5	.35	3.07	3.16
202.5	.18	2.05	1.44
207.5	.13	1.49	.92
212.5	.06	.70	.46
217.5	.02	.42	.30
222.5	.02	.23	.15
227.5	0	.14	.08
232.5		.14	.07
237.5		.14	.07
242.5		.05	.03
247.5		.05	0
252.5		0	0

Digitized by Google

The table on page 64 converts the same figures as just used to the "more than" basis, and adds an additional column for men 5 feet 8 inches tall, aged 35–39.

These figures are the basis for Diagram XIV. Diagram XV is the result of cumulating, on a "more than" basis, two of the theoretical curves of Diagram XI. Diagram XVI shows in the same way the age distribution of New York City policemen, firemen, and teachers, June 30, 1914.

Cumulative graphs with warped scale base. A difficulty is encountered when we try to use "more than" diagrams to compare

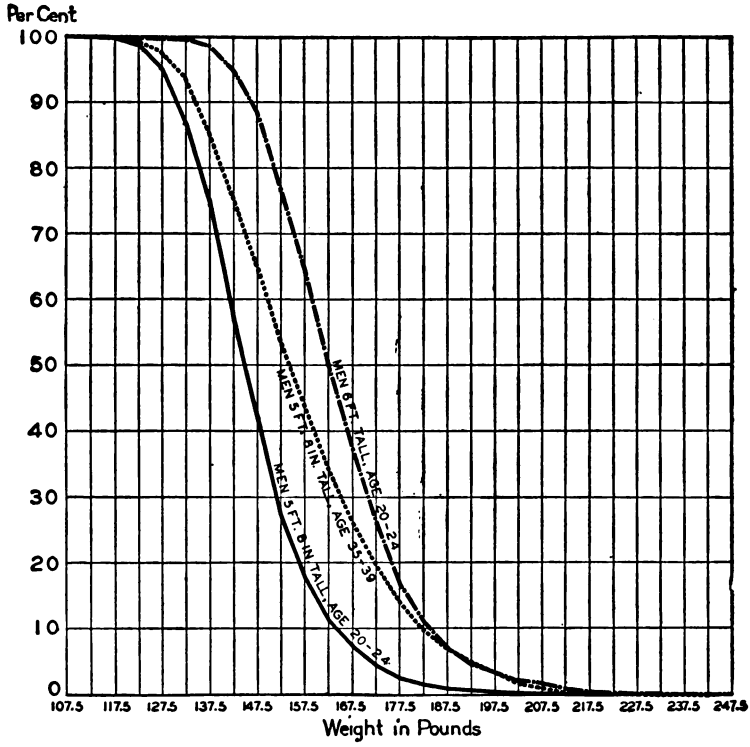


DIAGRAM XIV. PER CENT OF MEN HEAVIER THAN EACH WEIGHT INDICATED ON BASE LINE, FOR CERTAIN AGE AND HEIGHT CLASSES

Digitized by Google

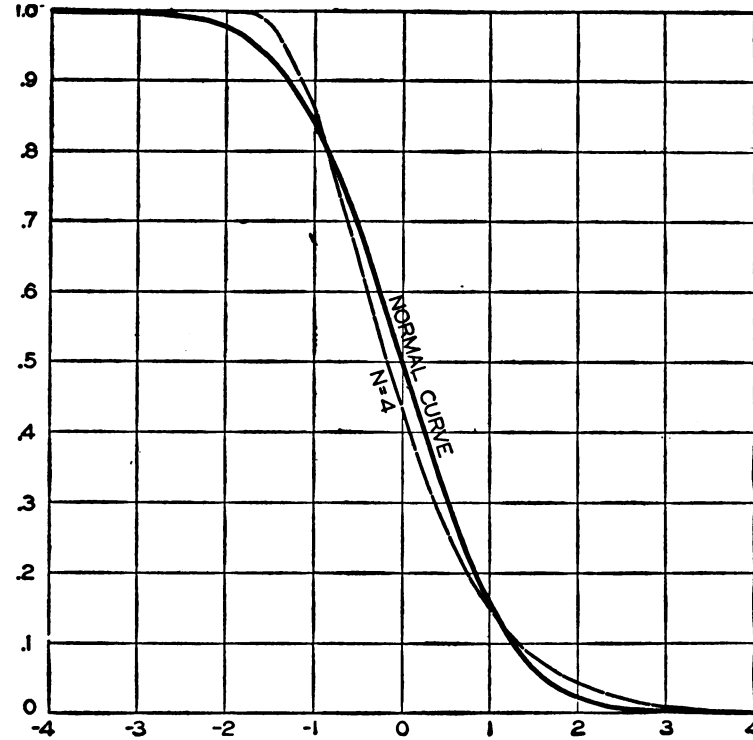


DIAGRAM XV. PROPORTION OF THE AREAS OF TWO THEORETICAL FREQUENCY CURVES LYING TO THE RIGHT OF VARIOUS ORDINATES

the distribution of incomes for several groups. It is characteristic of such distributions that there are many individuals with relatively small incomes and very few with large incomes; there are, however, in most groups a few persons with incomes 20 or 100 or even sometimes 500 times the average income. If the scale for incomes used on the base line is chosen so as to show the highest incomes, the graphs of the different distributions drop abruptly from the left hand upper corner and then run along close to the base line. Under such circumstances the comparison of two or more distributions is very poorly shown. If, on the other hand, a base scale is chosen which brings out the details for the lower incomes, the higher in-

Digitized by Google

comes are omitted entirely, unless the chart is made absurdly long. In order to meet this difficulty, the device of warping the scale along the base has been used with considerable success by the writer and his pupils.

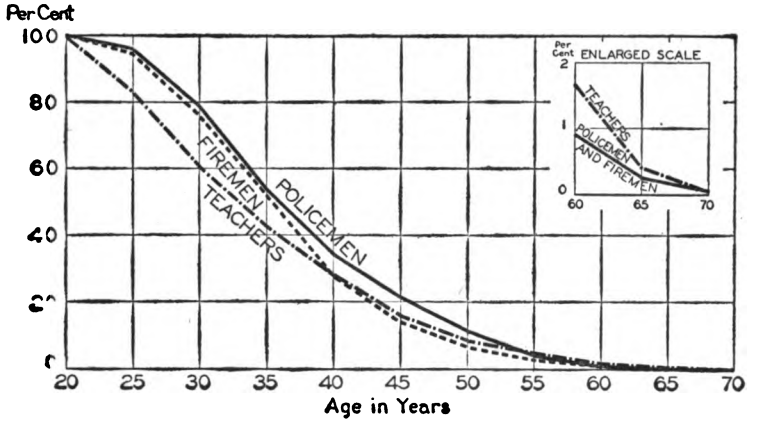


DIAGRAM XVI. PER CENT OF NEW YORK CITY (1) POLICEMEN, (2) FIREMEN, (3) TEACHERS, ON JUNE 30, 1914, WHOSE AGES EQUALLED OR EXCEEDED EACH AGE INDICATED ON BASE LINE

Specifically, in order to compare income distributions in different occupations, the 1916 income tax results for certain occupations were plotted on this warped scale, as shown in Diagram XVII. The method of warping is to space the vertical rulings of the grid in such a way that the "more than" graph of the distribution of all incomes is made to coincide with the straight line from the upper left-hand to the lower right-hand corner of the field. In detail, the spacing was done as follows: .

It was noted that 19.5 per cent of the incomes over $3000 fell in the $3000 to $4000 group, and 16.5 per cent in the $4000 to $5000 group. We have, then, that 80.5 per cent of incomes over $3000 exceeded $4000, and 64 per cent exceeded $5000. Let us then erect a vertical ruling at a point 80.5 per cent of the distance from the right-hand lower corner to the left-hand lower corner, and label that vertical "$4000." Similarly let us erect another ruling at a point

Digitized by Google

64 per cent of the distance from the right to the left corner and label that $5000. Similarly other verticals are located in accordance with the percentage of incomes exceeding $6000, $7000, and so forth. If the entire height of the diagram is taken as 100 per cent, it is geometrically obvious that the distance on the "$4000" vertical from the base to the left-to-right diagonal represents 80.5 per cent, and that the graph of "all incomes" is that diagonal. Using the grid thus constructed in exactly the same way as we use the ordinary unwarped "more than" grid, we can plot the cumulative percentages for various occupations.

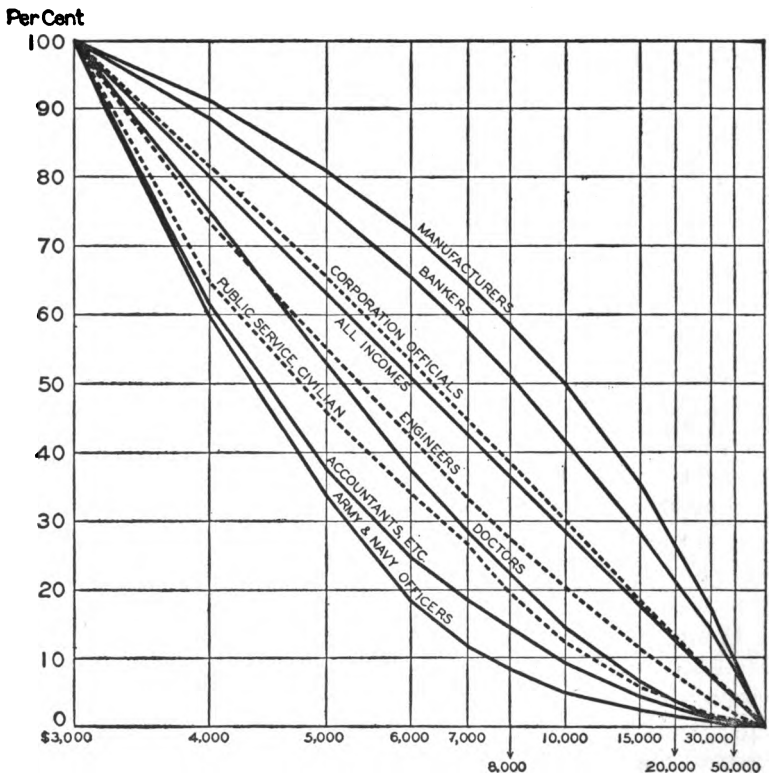


DIAGRAM XVII. COMPARATIVE DISTRIBUTION OF 1916 INCOMES IN CERTAIN OCCUPATIONS

Digitized by Google

Each curve in Diagram XVII shows for some occupation the percentage of persons making tax returns in 1916 whose net income exceeded the various sums indicated along the base line. No incomes of less than $3000 were reported; the curves therefore indicate the distribution of incomes above that mark rather than the relative standing of various occupations as a whole. Net incomes include income from property and investments as well as wages, salaries, and income from professions or business.[1]

While, of course, diagrams of this type are not valuable for presenting a single distribution, the scheme of warping the scale as indicated seems very promising for comparing two or more distributions. The warping of the scale is of course a complication, and should not be introduced if all the facts can be presented as clearly by the unwarped "more than" diagram.[2]

Graphical representation of quantity distributions. Quantity distributions are seldom analyzed except in connection with the corresponding frequency distribution. Consider, for example, the following table showing for each of the last four censuses the per cent of the population living in cities of each size group:

Population of Cities	1920	1910	1900	1890
1,000,000 or more	9.60	9.24	8.46	5.82
500,000 to 1,000,000	5.89	3.27	2.16	1.28
250,000 to 500,000	4.29	4.30	3.77	3.89
100,000 to 250,000	6.17	5.26	4.31	4.42
50,000 to 100,000	4.98	4.54	3.57	3.21
25,000 to 50,000	4.80	4.38	3.68	3.60
10,000 to 25,000	6.57	6.01	5.71	5.45
5,000 to 10,000	4.73	4.63	4.24	3.77
2,500 to 5,000	4.34	4.22	4.08	3.98
Rural	48.63	54.15	60.02	64.58

[1] From *Statistics of Income for 1916*, prepared by the Commissioner of Internal Revenue.

[2] Another way of warping the scales has been used to study and present income statistics. The method is based on taking the logarithms both of the incomes and of the number of persons receiving incomes in excess of various amounts. This method does not mean much except to mathematicians and engineers, and in some ways is deceptive. As recent investigations have shown that Pareto's law, with which it is connected, is not really valid, the method may well be left to the specialists. On this whole question the reader is referred to *Income in the United States*, published by the National Bureau of Economic Research.

Digitized by Google

In effect, we have here four quantity distributions, which we desire to compare. The value of the quantity varies in such a way, however, that we could not well plot either separate or cumulative figures directly. It turns out that a satisfactory graph is obtained by cumulating on the basis "living in cities of more than the population indicated," and basing the horizontal scale on the logarithms

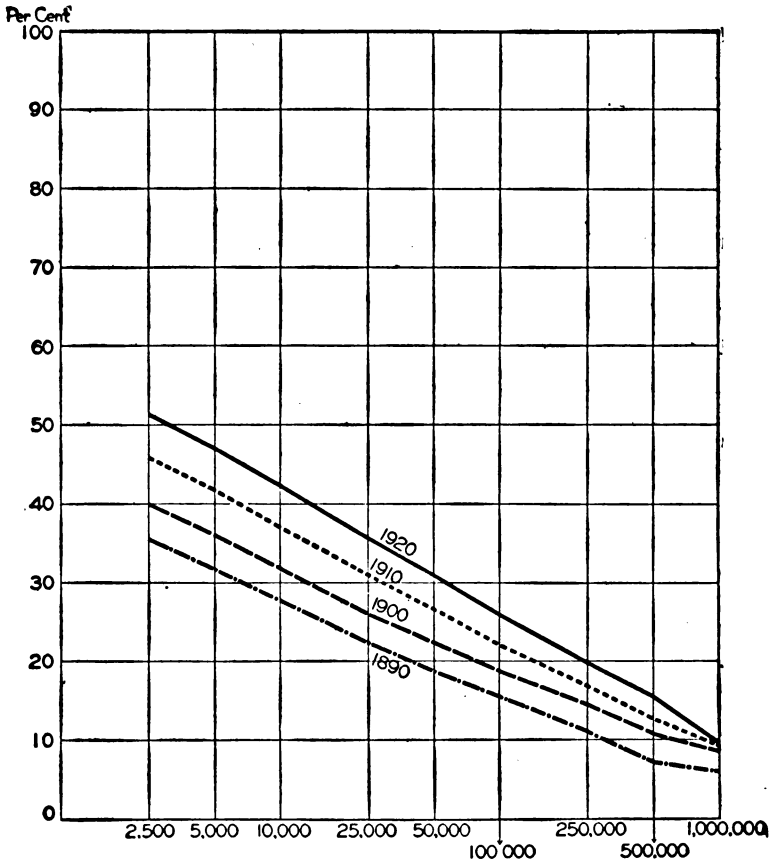


DIAGRAM XVIII. PER CENT OF THE POPULATION OF THE UNITED STATES LIVING IN CITIES WHOSE POPULATION EXCEEDED THAT INDICATED ON BASE LINE

Digitized by Google

of these populations. The accompanying diagram, drawn up on that basis, presents the facts as to the trend of population in a very compact form. The fact that these lines are so nearly straight suggests that we have thus discovered a law of population growth.

An interesting application of the warped scale "more than" diagram to two quantity distributions and one frequency distribution is given in Diagram XIX. This diagram shows the relation between number of taxed incomes, the income received by taxpayers, and the income tax paid.[1] The distribution of income received is taken as the standard for determining the position of the vertical rulings. The figures for 1920 incomes are given in the following table:

PER CENT OF NUMBER OF RETURNS MADE, OF AMOUNT OF INCOME RECEIVED, AND OF AMOUNT OF TAX PAID, BY THOSE WHOSE INCOME EXCEEDED SUM STATED

INCOME	RETURNS	INCOME	TAX
$1,000	100.00	100.00	100.00
2,000	63.20	82.94	96.57
3,000	27.81	56.88	92.34
4,000	15.49	43.96	88.14
5,000	9.39	35.65	84.57
6,000	6.95	31.57	82.46
8,000	4.37	26.16	78.74
10,000	3.11	22.73	75.47
15,000	1.68	17.46	68.87
25,000	.74	12.01	59.45
50,000	.21	6.50	45.10
100,000	.05	3.08	29.86
200,000	.0119	1.53	18.08

[1] Information similar to that given in Diagram XIX has more frequently been shown by a diagram suggested by Lorenz. As this diagram has not proved attractive except to some specialists, and is less generally applicable and less flexible than the warped scale diagram here suggested, fuller explanation here seems unnecessary. The reader is referred to King's *Elements of Statistical Method*, pp. 158–60, for an elementary explanation.

Digitized by Google

Which methods should be used for a given example? In order to tell which scheme of graphing frequency distributions will be best in a particular case, it is advisable to try all of them, at least roughly. The simple frequency polygon is most familiar, but in the opinion of the author the cumulative graphs are often more significant and will grow in popularity. In any case, remember that a clear title and explanation is about half the battle.

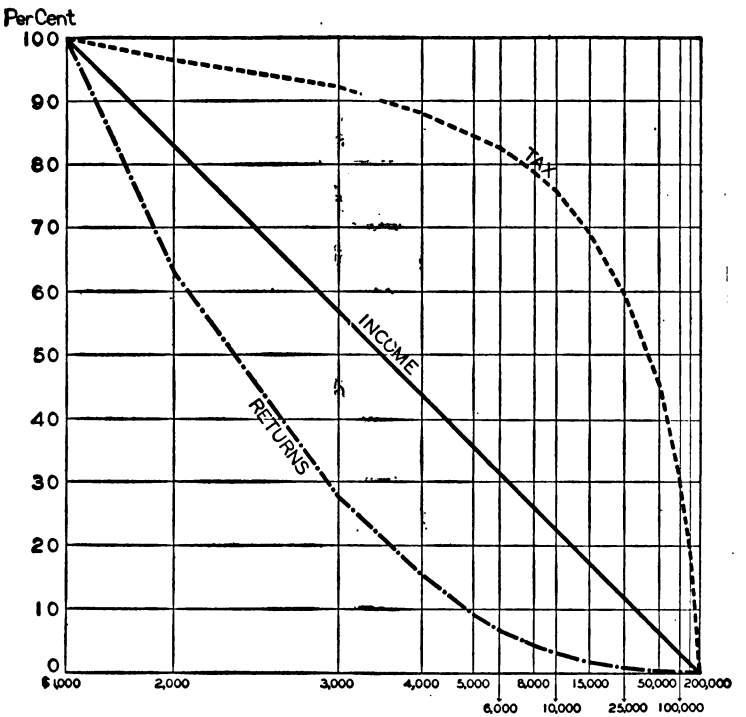


DIAGRAM XIX. PER CENT OF RETURNS MADE, PER CENT OF REPORTED INCOME RECEIVED, AND PER CENT OF TAX PAID, BY INDIVIDUALS WHOSE NET INCOME EXCEEDED AMOUNT STATED ON BASE LINE

Summary of Chapter IV. The next steps in statistical analysis, after the formation of statistical series and the construction of appropriate ratios, are concerned with particular types of series.

Digitized by Google

Methods used for historical series differ from those used for frequency distributions.

Frequency distributions, while not familiar by that name, really underlie the answers to many common questions.

The graphic representations of frequency distributions are essential to a real understanding of what they mean. Important graphical devices include:

(1) The histogram.

(2) The frequency polygon.

(3) The cumulative ("more than" or "less than") diagram.

(4) Cumulative diagrams with warped scale base.

Quantity distributions may be analyzed graphically by taking as ordinate of a curve the per cent of total quantity which belongs with values of the independent variable (population, income) in excess of a certain convenient boundary value.

A clear title and clear explanations are essential with all graphs of frequency distributions.

Digitized by Google

PROBLEMS

The following frequency distributions should be graphed in one or more of the ways discussed in the text. In each case the cumulative figures should be found and the percentage in each group, and the attempt made to grasp the facts summarized by the table, with special attention to which method (or methods) presents the situation most clearly.

1. The following table shows the number of men and women in a certain group who died in each age interval. The figures are from "A Genealogical Study of Population," by Carl E. Jones, *Quarterly Publications of the American Statistical Association*, December, 1918, vol. XVI, p. 207.

Ages at Death	Men	Women
18–22	0	1
23–27	3	6
28–32	5	9
33–37	15	13
38–42	27	21
43–47	23	26
48–52	19	23
53–57	34	26
58–62	35	27
63–67	48	39
68–72	49	44
73–77	52	34
78–82	54	44
83–87	43	24
88–92	20	21
93–97	11	7
98–102	3	1
103–107	1	1

Digitized by Google

2. The following table from the same source, shows the distribution of ages at marriage for substantially the same group as in problem 1:

Ages at Marriage	Husbands	Wives
13–17	2	99
18–22	303	732
23–27	691	461
28–32	282	120
33–37	91	37
38–42	40	14
43–47	15	7
48–52	9	1
53–57	3	0
58–62	4	1
63–67	1	0
68–72	1	0

3. The following table shows the number of American Rhodes Scholars of each age at time of matriculation at Oxford. The results are given separately for each year of matriculation at Oxford from 1904 to 1916. The figures are from "The Record of the American Rhodes Scholars," by the author, in the *American Oxonian*, January, 1921, vol. VIII, page 4.

Year of Matriculation	19	20	21	22	23	24
1904	1	4	9	8	7	2
1905	2	2	8	5	11	3
1907	1	4	13	8	8	3
1908	2	5	8	7	9	1
1910	1	8	7	7	9	7
1911	6	3	12	4	8	5
1913	2	4	11	9	4	7
1914	1	5	8	8	12	5
1916	2	1	2	8	3	4
Total 1904–16 inc.	18	36	78	64	71	37

Digitized by Google

4. The following table is based on a study of the meal checks at a cafeteria in Evanston, Illinois. The table is quoted from an article by Horace Secrist, in the *Quarterly Publications of the American Statistical Association*, vol. XVI, September, 1919, p. 471.

NUMBER OF MEN (OR WOMEN) WHO PAID FOR BREAKFAST, DINNER, OR SUPPER THE SUMS INDICATED

AMOUNT PAID FOR A MEAL (in cents)	ALL MEALS			BREAKFAST			DINNER			SUPPER		
	Total	Men	Women	Total	Men	Women	Total	Men	Women	Total	Men	Women
3 to 7	15	7	8	5	1	4	6	2	4	4	4	
8 to 12	188	64	124	84	25	59	57	12	45	47	27	20
13 to 17	516	150	366	252	91	161	183	39	144	81	20	61
18 to 22	763	230	533	220	87	133	356	98	258	187	45	142
23 to 27	982	343	639	134	65	69	552	186	366	296	92	204
28 to 32	849	345	504	70	42	28	497	211	286	282	92	190
33 to 37	672	315	357	34	25	9	350	179	171	288	111	177
38 to 42	758	334	424	19	11	8	336	174	162	403	149	254
43 to 47	702	351	351	14	12	2	297	164	133	391	175	216
48 to 52	563	307	256	2		2	224	120	104	337	187	150
53 to 57	407	223	184	1		1	159	89	70	247	134	113
58 to 62	179	106	73				77	47	30	102	59	43
63 to 67	100	49	51				60	29	31	40	20	20
68 to 72	75	37	38				38	23	15	37	14	23
73 to 77	36	16	20				18	7	11	18	9	9
78 to 82	17	11	6				9	5	4	8	6	2
83 and over	21	9	12	1		1	14	6	8	6	3	3

Digitized by Google

5. The following table gives the number of contributors to Charitable organizations in Providence who made the number of contributions indicated on the left, and also corresponding figures for that portion of these contributors whose total of contributions was less than $20.00. The figures are from *Statistical Survey of Contributions to Providence Voluntary Community Organizations*, by the author.

Number of Contributions	Contributors Giving Indicated Number of Times	Contributors whose Total of Contributions was Less than $20.00
1	3806	3393
2	855	593
3	398	202
4	230	69
5	150	23
6	89	6
7	77	4
8	47	1
9	39	0
10	33	0
11	18	0
12	19	0
13	8	0
14	4	0
15	9	0
16	3	0
17	5	0
19	1	0
21	3	0

Digitized by Google

6. The following table gives in round numbers the number of men and women of each height, according to the Medico-Actuarial investigation, based on the records of physical examination of applicants for insurance. Each figure in the table represents 100 individuals.

NUMBER OF MEN AND OF WOMEN OF EACH HEIGHT IN A TYPICAL GROUP

HEIGHT	MEN	WOMEN
4 ft. 8		1
4 ft. 9		2
4 ft. 10		6
4 ft. 11	1	13
5 ft.	4	55
5 ft. 1	5	75
5 ft. 2	12	147
5 ft. 3	26	184
5 ft. 4	66	251
5 ft. 5	121	228
5 ft. 6	221	182
5 ft. 7	281	114
5 ft. 8	375	66
5 ft. 9	322	25
5 ft. 10	327	11
5 ft. 11	230	4
6 ft.	146	1
6 ft. 1	48	
6 ft. 2	23	
6 ft. 3	7	
6 ft. 4	2	
6 ft. 5	1	
	2218	1365

Digitized by Google

7. The following table shows the number of high school teachers in certain cities who in 1918–19 received the salaries indicated. The first group of cities, marked "I," includes cities of over 100,000 population; the second marked "II," those of 25,000 to 100,000 population; the third (III) those of 10,000 to 25,000 population; the fourth (IV) those of less than 10,000 population. The figures are summarized from the Report on Teachers' Salaries and Salary Schedules, prepared by E. S. Evenden for the Commission on the Emergency in Education of the National Education Association.

Salary	I	II	III	IV	Total
$300– 399	2	2	0	2	6
400– 499	3	5	0	3	11
500– 599	2	9	6	9	26
600– 699	9	19	26	63	117
700– 799	34	64	97	198	393
800– 899	49	121	155	424	749
900– 999	84	243	254	432	1013
1000–1099	468	294	240	234	1236
1100–1199	182	244	131	88	645
1200–1299	268	385	137	106	896
1300–1399	213	248	128	72	661
1400–1499	194	268	112	33	607
1500–1599	306	170	105	62	643
1600–1699	237	268	58	22	585
1700–1799	215	61	32	5	313
1800–1899	355	80	27	7	469
1900–1999	250	34	6	2	292
2000–2099	90	20	20	3	133
2100–2199	61	17	4	1	83
2200–2299	23	14	7	0	44
2300–2399	14	4	0	0	18
2400–2499	35	1	1	0	37
2500–2999	0	10	1	0	11
3000 and over	0	2	0	0	2
All salaries	3094	2583	1547	1766	8990

Additional frequency distributions to which the methods of this chapter may be applied will be found in Yule's *Introduction to the Theory of Statistics*, pp. 88, 90, 95, 159, 161; in the problems at the end of Chapter X (correlation) in this book, and in the *World Almanac*, 1924, p. 311, "Number of Taxable Incomes, by Classes." These problems in each case involve two or more distributions which are to be compared; after the student has worked over a number of such cases, he will be familiar with the characteristic features of frequency distributions, and on inspection of the frequency polygon for a single distribution will be able to grasp its peculiarities.

Digitized by Google

CHAPTER V

AVERAGES AND THEIR USE

ONE average, the arithmetic mean, familiar to almost every one as "the average," was discussed briefly in Chapter II. In that chapter, it was regarded merely as a ratio — the ratio of two related quantities, each of which was a total. From the total number of ton-miles of traffic on a certain railroad in a certain time, and the total cost of moving that traffic, we easily derive the cost per ton-mile. Similarly we find amount of stock per stock-holder, yield of corn per acre, and other frequently used "size-ratios."

From another point of view, however, the arithmetic mean may be regarded, not merely as a ratio of two totals, but as a typical number selected to represent the magnitude of the quantity in a group of individuals. Or it may be regarded as a summarizing figure for a frequency distribution, giving the location of the central position. Its significance from this point of view lies largely in the fact that frequency curves are usually somewhat bell-shaped, and that therefore a number corresponding to a point near the center differs little from most of the other numbers included in the group.

From this broader point of view, it is at once clear that there may be other typical numbers associated with a frequency distribution, in addition to the arithmetic mean. Such numbers are termed *averages*, and will be discussed in the present chapter.

What an average does not do. No user of averages should forget, however, that the average does not represent the value of the quantity for every individual in the group. The complete frequency distribution is often needed as a basis for wise action or sound judgment. As an instance of over-emphasis on the average, consider the following editorial [1] concerning the average number of persons per bath-tub in Boston. Noting that the figures show an average of one bath-tub for every 4.4 persons, the writer continues:

[1] From the *Providence Journal*, March 19, 1921.

Digitized by Google

"Taking the 'average family of five,' an invention of the statistician, it appears that there are already more than enough to supply a tub to every Boston family. One to every inhabitant would seem extravagant, considering the business depression, though, to be sure, the ratio would not necessarily, or probably, mean that every inhabitant had a tub.

"That's the trouble with statistics; they rarely mean what they seem to. At the present time not every 4.4 persons have a tub, though that is the average quota."

The point is well taken if we say, not "trouble with statistics," but "trouble with averages." The editorial writer tries to philosophize on information which gives the general aspect of affairs, not realizing that for the analysis he is trying to make, he needs more detailed statistics — to be specific, a frequency distribution.

Definition of the arithmetic mean, and elementary formulas. The arithmetic mean (A.M.), usually called "the average," may be defined as the sum of the measures of a certain characteristic for a group of individuals divided by the number of individuals. For instance, if the characteristic is age, the average age of three men aged 30, 32, and 40, respectively, is 30 + 32 + 40 divided by 3, or 34. In symbols, if $x_1, x_2 \ldots x_n$ denote the measures for the first, second, and so forth up to the n'th individual, and $\bar{x}$ denotes the average, the formula is:

$\bar{x} = \dfrac{x_1 + x_2 + \ldots + x_n}{n} = \dfrac{\Sigma x}{n}$, where Σ indicates that the sum of all quantities x, no matter what the subscript, is to be taken.

In applying the above definition, and using the formula, it often happens that many of the measures are equal, at least to the degree of accuracy that seems significant in the discussion. In such cases, the formula may be modified in appearance, though not in fact. Instead of writing $\bar{x} = \dfrac{x_1 + x_2 + \ldots + x_n}{n}$, if the first three x's are the same, we may write $\bar{x} = (3x_1 + x_4 + \ldots + x_n)/n$. In general, if the measure x_1 applies to f_1 individuals, x_2 to f_2, and so forth, the formula becomes:

Digitized by Google

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \ldots + f_nx_n}{f_1 + f_2 + \ldots + f_n} = \frac{\Sigma fx}{\Sigma f}$$

If no item occurs more than once, so that the f's are all unity, this formula reduces to the form first given, and the resulting average is sometimes called the "unweighted mean." In our opinion, however, the phrase "mean with equal weights" is preferable, as the formula should be applied only when the question of weights has been considered and a decision reached that all the available evidence indicates that every x occurs with the same frequency. The phrase "unweighted" seems to imply that no decision has been reached about what weights are appropriate, in which case no average whatever is justified.[1]

A useful modification of the formula for the weighted average is obtained by dividing each f by the total of the f's before multiplying. Such weights, multiplied by 100, are referred to as *percentage* frequencies or *percentage* weights. If we denote by w_1 the quotient f_1 divided by the sum of the f's, we clearly have

$$w_1 + w_2 + \ldots + w_n = \frac{f_1}{f_1 + f_2 + \ldots + f_n} + \frac{f_2}{f_1 + \ldots + f_n} + \ldots + \frac{f_n}{f_1 + \ldots + f_n} = \frac{f_1 + f_2 + \ldots + f_n}{f_1 + f_2 + \ldots + f_n} = 1; \text{ and}$$

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \ldots + f_nx_n}{f_1 + f_1 + \ldots + f_n} = \frac{f_1}{f_1 + \ldots + f_n}x_1 + \frac{f_2}{f_1 + f_2 + \ldots + f_n}x_2 + \ldots + \frac{f_n}{f_1 + \ldots + f_n}x_n = w_1x_1 + w_2x_2 + \ldots + w_nx_n.$$

The use of both the original and the modified formulas is illustrated in the following example:

Find the average weekly earnings of a group of ten men, two of

[1] The use of the term "weights" rather than "frequencies" will be obviously appropriate to any one who is familiar with the determination of center of gravity in mechanics. The transfer of this term from mechanics to statistics took place by way of the theory of errors. When the arithmetic mean of a series of measures of a physical or astronomical magnitude is to be determined, some measures are thought to be more reliable than others and therefore counted 2, 3 or more times as compared with once for the less reliable observations. That is to say, the reliable observations are assigned a virtual frequency greater than their actual frequency. To the physicists and astronomers, it seems natural to call such virtual frequencies "weights."

Digitized by Google

whom receive \$20 per week, five of whom receive \$25, and three of whom receive \$30.

By direct substitution in the first formula,

$$\bar{x} = \frac{2 \times 20 + 5 \times 25 + 3 \times 30}{2 + 5 + 3} = \frac{255}{10} = 25.5$$

In the second formula, $w_1 = \frac{2}{10} = .2$, $w_2 = \frac{5}{10} = .5$, and $w_3 = .3$; and hence $\bar{x} = .2 \times 20 + .5 \times 25 + .3 \times 30 = 25.5$.

It will be found in practice that the second formula is often more convenient and that the steps in the computation are more easily checked. The weights are, moreover, in a more convenient form to quote. The formula shows that what is significant in the weights is not the frequencies singly, but the proportion which each frequency is of the total. If therefore we can determine or estimate these proportions without getting at the actual frequencies, the results will be as satisfactory as if we had determined or estimated the frequencies themselves.

Computation by deviation method. The computation of the A.M. is often simplified, if the number of cases is large, by assuming a position for the mean, and computing a correction c as follows: Let x_M be the assumed mean, $\bar{x}$ the true mean.

$$\begin{aligned} c = \bar{x} - x_M &= w_1x_1 + w_2x_2 + \ldots + w_nx_n - (w_1 + w_2 + \ldots + w_n)\, x_M \\ &= w_1(x_1 - x_M) + w_2(x_2 - x_M) + \ldots + w_n(x_n - x_M) \\ &= \frac{f_1(x_1 - x_M) + f_2(x_2 - x_M) + \ldots + f_n(x_n - x_M)}{f_1 + f_2 + \ldots + f_n} \end{aligned}$$

That is to say, the correction is computed by subtracting the assumed mean from each value of the variable, multiplying each difference, or "deviation," by the appropriate frequency, and dividing the algebraic sum of these products by the total of the frequencies. An application of this method to a numerical case follows. In this example, as is always done with frequency distributions when the corresponding quantity distribution is not available, the assumption is made that without significant error all items in a frequency class may be taken as concentrated at the mid-point of the class interval, and the classification "4 inches," for instance, means 3.5–4.5 inches.

Digitized by Google

ARITHMETIC MEAN OF THE LENGTH OF 326 EARS OF CORN
(Computation by Deviation Method)

LENGTH	FREQUENCY — NO. OF EARS	DEVIATION FROM ASSUMED MEAN 9	DEVIATION TIMES FREQUENCY
3 in.	1	−6	−6
4	1	−5	−5
5	3	−4	−12
6	14	−3	−42
7	25	−2	−50
8	62	−1	−62
9	119	0	0
10	80	1	80
11	19	2	38
12	2	3	6
Total	326		124–177

$$c = \frac{124 - 177}{326} = -\frac{53}{326} = -.16; \quad \bar{x} = x_M + c = 9 - .16 = 8.84.$$

Effect of a change in weights. The effect of a change in the proportionate weights is the same as that discussed under percentage analysis in Chapter II. If w_i and w_i' are two sets of weights applied to the same x's, the two means are

$$\bar{x} = w_1x_1 + w_2x_2 + \ldots + w_nx_n$$
$$\bar{x}' = w'_1x_1 + w'_2x_2 + \ldots + w'_nx_n$$
$$\bar{x} - \bar{x}' = (w_1 - w'_1)x_1 + (w_2 - w'_2)x_2 + \ldots + (w_n - w'_n)x_n.$$

If the difference of weights is random, and the number of terms is not too small, the + and − terms on the right side tend to cancel; if on the whole, however, the second set of weights has meant a greater stressing of the large x's and a smaller stressing of the small x's, the right hand side will be −, and the second weighted mean is larger than the first. And vice versa, if the second set of weights has shifted emphasis from the *large* to the small x's, the right hand side will be +, and the second weighted mean smaller than the first.

The importance of such changes of weights is illustrated by the following example (Example 1), which shows that average rates of

Digitized by Google

pay by occupations do not permit immediate inferences as to general averages.

EXAMPLES FOR STUDENT

1. Two stores employed laborers at an average pay of \$3.00 per day, salesmen at \$5.00 per day, bookkeepers at \$6.00 per day, and executives at \$10.00 per day. The percentages of employees in the different classes were, respectively, 10, 50, 30, 10, and 15, 55, 25, 5. State on inspection which store pays the highest average wage, and find both averages.

2. Modify the example so that the rates paid by one store will be larger for each type of employee, while the general average rate will be smaller, than for the second store.

3. Show by an example that the average size of a family, counting each family once, is smaller than the average size of family, counting each family once for each child.

The median. Another number often used to advantage in summarizing a frequency distribution is the *median*, defined as the magnitude of the middle item in a series arranged as a rank list. If the number of items in the series is even, the average value of the two middle items is usually taken as the median, although it is often clearer to say that one half of the values are a or less and one half $a + 1$ or greater.

Two examples will illustrate this definition in the discrete case and point out an incorrect application of the definition. Two classes of children are given a test in arithmetic, to do in a limited time; and they are scored on the basis of number of problems done correctly; with the following results:

No. of Problems Correct	Children with Score Indicated	
	1st class	2nd class
0	0	1
1	4	4
2	3	7
3	13	5
4	10	6
5	5	1
6	1	
7		
	36	24

In the first class, the median should be the *average* of the number of

Digitized by Google

problems solved by the 18th and 19th child; as both of these children are included in the group who solved 3 problems, there is no difficulty — the median is 3. In the second class, the median falls halfway between the performance of the 12th and 13th child, i.e., halfway between 2 and 3, and may be called $2\frac{1}{2}$. A better way is to say that half the class did 2 or less, and half 3 or more, thus summarizing the frequency distribution without artificiality.

Some writers use a process of interpolation on these problems in the way that will be explained in the next paragraph for continuous frequency distributions and obtain for answer $3\frac{1}{18}$ and 3 respectively, on the theory that the children who did three problems were distributed over the interval between 3 and 4. But as the conditions of the grading as originally stated expressly excluded fractional values, it seems artificial to get fractions into the answer by an untested assumption. If this method is used in cases such as this, in which the original data do not admit fractions, or in which it is known that the individual items are not uniformly distributed over the interval, it is suggested that the result be known as the pseudo-median to distinguish it from the real median, the value of the variable quantity for the middle item of the ranked series.

There are, of course, a large number of cases in which the individuals are, as far as is known, uniformly distributed over the interval, and the process of interpolation is required. It may be noted in passing that if it is possible to go back to the original data and find the value of the characteristic for the middle individual, exactly, the result is, strictly speaking, more accurate than the following. But such a proceeding is seldom worth while, if the data are given in the form of a fairly detailed frequency table.

Let us use the example for which the arithmetic mean was computed in the preceding section:

Length of ears (inches)	3	4	5	6	7	8	9	10	11	12
No. of ears	1	1	3	14	25	62	119	80	19	2
Cumulative no.	1	2	5	19	44	106	225	305	324	326

By definition of median, it is the length of the ear halfway between the 163d and the 164th. Both of these ears are in the 9 in. (i.e.,

Digitized by Google

8.5 to 9.5) group. The natural assumption as to distribution in the interval is that the lengths of the 119 ears are equally spaced from 8.5 to 9.5. We may think of this interval as divided into 119 small equal sub-intervals, with one ear assigned to each sub-interval, and with the length of the ear given exactly by the mid-point of the sub-interval. That is to say, the lengths of successive ears are $8.5 + \frac{1}{238}$, $8.5 + \frac{3}{238}$, etc. To reach the 163d and 164th ear in the whole series, we must go on to the 57th and 58th, respectively, in this inch interval from 8.5 to 9.5 (163 − 106 = 57). The length of the 57th ear in this interval is $8.5 + \frac{113}{238}$, and of the 58th $8.5 + \frac{115}{238}$, the point half way between is $8.5 + \frac{114}{238} = 8.5 + \frac{57}{119} = 8.979$. I.e., if n denotes the total number of items, Median = beginning of interval $+ \frac{[\frac{n}{2} - (\text{number of items before interval})] \times (\text{width of interval})}{\text{number of items in interval}}$.

If the length of the ear is thought of as given by the *beginning* of the sub-interval in which it lies, the result should be $\frac{1}{238}$ less; if it is thought of as given by the *end* of its sub-interval, the result should be $\frac{1}{238}$ more. The latter result is often obtained as the result of a blind use of the formula applicable to discrete series, that the median is the value of the character for the $\frac{n+1}{2}$ th individual; but the scheme outlined above, i.e., using $\frac{n}{2}$ rather than $\frac{n+1}{2}$, is recommended because it works both ways; the same value is found by starting at the higher end of the series and working down as was just found by beginning at the bottom and working up.

The *median* should be used, either alone or with the mean, when it gives or helps to give a good summary of the frequency distribution. It is most valuable when it differs from the mean; if it is

Digitized by Google

nearly the same as the mean, the latter is the better single summary because of the greater general familiarity with it. When they are different, however, the median is usually the more typical figure — the one which it would be better for people to bear in mind as about the value of the character for the entire group. For instance, suppose that in a group of 10 men, 4 have incomes of $2000, 3 of $3000, 1 of $4000, 1 of $5000 and 1 of $10,000. The median income is $3000, while the mean is $3600; it will be noted that only three individuals have incomes in excess of the mean, while 7 are below. In general, the median is better than the mean for wages, or incomes, or estates, because it is less influenced by a few large exceptional cases.

The general principle may be stated thus: Find the median, and consider it in connection with the frequency distribution which it tries to summarize. If it seems as typical as a single figure can be, use it. It is often advantageous to state in words just what the median means — e.g., in our ears of corn example, half the ears are longer than 8.98 inches and half shorter. Or more roughly, 9 inches is the half way point in length of ear. Thus the statistician may postpone justifying the imputation that he is "high brow" until he deserves it, perhaps, in connection with the more elaborate analysis that we shall discuss later.

The mode. The *mode* is the most usual or frequent value of the character, that for which the instances are most numerous. For discrete distributions, it is easily and accurately determined by inspection — for the two classes at the beginning of the preceding section, the modes are 3 problems and 2 problems, respectively. In the frequency distribution of lengths of ears of corn, the example introduced on page 86, the mode is 9 inches; but this means the class interval 8.5 to 9.5, and cannot at this stage of the work be located more accurately. If a different classification is adopted — say 8–9, 9–10, etc., the mode would be shifted; if the attempt is made to locate the mode more exactly by grouping, say, in quarter inch intervals, the chance variations in the grouping are likely to become too prominent to permit a trustworthy determination of the largest class.

Digitized by Google

The chief advantage of the mode is the fact that it is a method commonly used in ordinary speech. The Bible says that "the years of a man's life are three score years and ten"; we say that the first killing frost usually comes about September 20th, or that young men of a certain race or community usually marry at 28 to 30, etc. The mode is essentially just this sort of thing; and the impossibility of making it exact does not detract greatly from its usefulness. If an exact figure is wanted, it is better to use the mean or median. The mode should never be used in discussing a small number of cases, where "usually" would be an inappropriate adverb.

Relation of mode to mean and median. For a certain type of frequency curves, which will be discussed later, the following relation is very nearly true: 3 (Mean-Median) = Mean-Mode. As this is roughly true of many actual frequency distributions, it may be used to assign an exact position to the mode; if found in this way, however, the mode does not give us new information about the frequency distribution, but merely states in another form what is already implied by the mean and the median.

Illustration of mean, median, and mode from the mortality table. The distinction between mean, median, and mode is well illustrated from a mortality table. Consider, for instance, in the table of native white males in the *United States Life Tables*, 1910, the figures for men of exact age 21. The mean length of life for persons reaching any age is given by the table as the "complete expectation of life." For age 21, the mean period of survival is therefore 42.52 years, as stated in the last column. The median period of survival is that from age 21 to the age at which half of those then alive will be dead; at 21, 7854 is the number of survivors, as shown in the first column; 3927, one half of this, occurs between exact age 66 and 67, and by interpolation is 66.97 (let the student check this). The median period of survival is 45.97. The mode is the commonest period of survival; as more men die between exact ages 75 and 76, than between any other two consecutive ages, 75 is the modal age at death, and 54 therefore the modal period of survival. We have therefore Mean 42.5, Median 46.0–, and mode 54.

Digitized by Google

The harmonic mean. Certain problems require the use of an unfamiliar average, the harmonic mean, defined as the reciprocal of the arithmetic mean of the reciprocals of the given quantities. The algebraic formula is

$$\frac{1}{H} = \frac{\frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{a_3} + \ldots + \frac{1}{a_n}}{n}$$

Its significance will be brought out by the discussion of an example which actually arose in more complicated form in a statistical office in Washington during the World War.

Suppose 20 boats make six transatlantic trips each per year, giving as the time for a "turn around" (i.e., time between consecutive departures from the same ports) one sixth year — approximately 60 days, and that 10 boats make 4 trips each per year, giving as their time for a "turn around" one fourth year, approximately 90 days. (A year of 360 days is used merely for convenience.) What is the average number of days per turn around?

An obvious method of determining the average number of days per trip would be to weight 60 by 20, the number of boats requiring 60 days, and 90 by 10, add the products, and divide by 30, the total number of boats. This computation gives 70 as the average number of days per trip. (Method I.)

An equally obvious method is to find first the average number of trips per year per boat, as follows:

Total trips $= 20 \times 6 + 10 \times 4 = 120 + 40 = 160.$

Average per boat $= \frac{160}{30} = 5\frac{1}{3}$. (Method II.)

But a boat which makes $5\frac{1}{3}$ trips per year (360 days) makes one trip in 67.5 days rather than in 70 days, as Method I would imply.

It is obvious that one or the other of these "obvious" methods is wrong. Closer inspection reveals no flaw in the second method, but suggests the possibility of an error in the weights used in Method I. The product of number of days per trip and number of boats requiring that number of days, is not total number of days required in a year by boats of that class. We should multiply also by number of trips per year made by each boat of that class. In effect, then, the weight to apply to the turn around of each class of boats is the

Digitized by Google

total number of trips accomplished in a year by boats of that class — that is to say, the number of boats in that class multiplied by number of trips per boat per year.

If we revise the weights along these lines, the arithmetic of Method I becomes:

$$\text{Av. no. of days} = \frac{60 \times 20 \times 6 + 90 \times 10 \times 4}{20 \times 6 + 10 \times 4} = 67.5 \text{ (Method I–A)}$$

We may now proceed to modify the arithmetic in a few details so as to lead to a slightly different equation, which we shall find can be set up directly from the original statement of the problem.

We note, then, that each product in the numerator of the preceding fraction really contains 360 as a factor. If we divide both numerator and denominator of the fraction by 360, we have:

$$\text{Av. no. of days} = \frac{20 + 10}{20 \times \frac{1}{60} + 10 \times \frac{1}{90}}$$

Or we may say:

$$\frac{1}{\text{Av. no. of days}} = \frac{20 \times \frac{1}{60} + 10 \times \frac{1}{90}}{20 + 10}$$

We now notice that the correct average number of days may be obtained, as this formula indicates, by finding the harmonic mean of number of days per trip for the different types of boat, using the number of boats of a given type as the weight to be applied to the reciprocal of the corresponding "number of days per trip."

We see, then, that correct results may be obtained in this problem by any one of the following three methods:

METHOD II. Find the arithmetic mean of trips per year for the various classes of boats, weighted in each case by number of boats in that class. The average number of days per trip is then found by dividing number of days in a year by average number of trips per year.

METHOD I–A. Find the arithmetic mean of days per trip for the various classes of boats, weighted in each case by

Digitized by Google

number of trips per year accomplished by boats of that class.

METHOD III. Find the harmonic mean of days per trip for the various classes of boats, weighted in each case by number of boats of that class.

As Method I–A is relatively awkward, the real choice lies between Method II, "A.M. of trips," and Method III, "H.M. of turn arounds." As Method II is based on the familiar arithmetic mean, it seems preferable on the general principle that the simplest technique is best, unless it is inaccurate or for some special reason less desirable. When the problem arose in Washington, however, the interest of the executives in charge of overseas transportation was concentrated so directly on "turn around," as the readiest measure of operating efficiency, that it seemed more straightforward to avoid introducing "trips per year" into the computation and to use Method III rather than Method II.

It may be mentioned that the danger of seeming abstract and technical was in large measure overcome by describing the result of the averaging by Method III as the "effective average" rather than the harmonic mean. Under this name it was accepted and used as a valid and significant measure of performance. The moral is: use as exact and careful statistical technique as is necessary to secure reliable results, but present your results to the practical man in familiar phraseology.

Additional illustrations of the use of the harmonic mean. Let us try to work toward a general statement of the reasons for using the harmonic mean of the outstanding quantities in a discussion rather than their arithmetic mean. Careful analysis of the preceding problem showed that the quantity which should be averaged by the arithmetic mean process was number of trips per year, essentially the reciprocal of number of years (or days) per trip, the outstanding quantity from the point of view of those who were dealing with the problem. In all such cases, in which the quantity whose arithmetic mean should be taken is the reciprocal of the "outstanding" quantity, we may secure correct results by finding either the arithmetic mean of the different values of the first quan-

Digitized by Google

tity or the harmonic mean of the different values of the "outstanding" quantity.

The difficulty, then, reduces to the logical difficulty of deciding which quantity "should" be averaged. The answer to this question depends on the circumstances of each particular case, and sometimes on what the resulting average is to be used for when found. If the purpose is to find an average which can be used in a second computation, the quantity which should be averaged is the one whose average gives the same result in the second computation as the detailed data would give. If the purpose is less definite, the choice must be made on somewhat less definite grounds. In any case, it is of great assistance to be familiar with typical instances in which the choice had to be made between averaging a quantity or its reciprocal. We shall therefore discuss here, or suggest for the student's discussion, several such instances, and then return to the question of how the choice should be made when it is to be made "on somewhat less definite grounds."

Such a typical instance arises in the purchase or sale of British funds in the New York foreign exchange market. Suppose two such purchases are made, one at the rate of $4 per pound, and the second at the rate of $5 per pound. Or looking at it the other way, we may quote the rates as 5 shillings for a dollar and 4 shillings for a dollar. The arithmetic mean of the rates as first stated is $4.50 per pound, whereas the arithmetic mean of the rates stated in the second form is 4.5 shillings per dollar, equivalent to $4.44 per pound. As the student may readily verify, the average of $4.50 per pound is correct if the same number of pounds is purchased at each rate. On the other hand, if the same number of dollars is invested at each rate, the correct average is 4.5 shillings per dollar, or $4.44 per pound. If actual purchases are not being made or contemplated by the person for whom the average is found, the choice of average will depend on whether it seems more natural to regard the number of pounds or the number of dollars as the fixed element in the several transactions averaged. Probably in most cases the New York purchaser will regard the pound as the fixed unit whose price varies, and will therefore average dollars per pound. Similarly the London purchaser is more likely to take the other point of view. As these

Digitized by Google

points of view are actually exemplified in the manner of quoting the rates on the two markets, the weekly average based on the London rates will not correspond exactly with the weekly average based on the New York rates, even if the daily rates correspond exactly (unless, of course, there is no fluctuation during the week). To be sure, unless the changes during the week are of unusual size, the difference between the two averages will not be significant, and will ordinarily be concealed by the fact that official daily quotations on the two markets differ slightly for other reasons.

It should be noted that weighting the rates by the actual volume of transactions would remove the discrepancy noted, provided that the same transactions constituted the basis for the weights used in both averages.

EXAMPLES FOR STUDENT

1. Suppose that the prices of wheat and corn at two dates are as given in the following table:

	July 1, 1928	July 1, 1929
Wheat	.80	1.00
Corn	1.00	.50

Show that the arithmetic mean of the relative numbers for the second year, first year as base, is the reciprocal of the harmonic mean of the relative numbers for the first year, second year taken as base. (Relative numbers are regarded as ratios rather than percentages in this problem.) In the next chapter, on "Index Numbers," further reference will be made to this principle.

2. A dealer in eggs regularly sold each day 10 dozen eggs at 12 for a dollar and another 10 dozen eggs at 20 for a dollar. One day he decided to simplify matters by selling his entire daily quota at the average price, as he figured it, of 16 for a dollar. At the end of the day, however, he was surprised and grieved to find that he had taken in only $15 as compared with the usual $16. His surprise was increased when he talked with another egg dealer, who had tried the same scheme and found that it worked satisfactorily. The daily quota of the second dealer however was 10 dozen of the 12 for a dollar eggs and 200 of the 20 for a dollar eggs.

Explain the reason for the lost dollar, and state what average price the first dealer should ask to secure his customary receipts.

Which ratio should be averaged? — General principles. The problems discussed or suggested above are typical of those which arise when ratios are to be averaged. The arithmetic mean of

Digitized by Google

trips per year is correct, rather than the arithmetic mean of days per trip, when the total number of trips in a fixed interval of time is to be deduced from the result. The arithmetic mean of prices in dollars per pound is correct if the number of pounds purchased at various prices is the same. The arithmetic mean of prices stated in the form, so many cents for a fixed unit or a fixed number of units, is correct if the total amount paid for a fixed list of items is supposed to be consistent with the resulting average. If on the other hand the amount of money spent is fixed and the number of items secured is to be deduced from the average, the correct average to use is the arithmetic mean of prices stated in the form, so many units for a dollar or other fixed sum of money.

In general terms, then, the principle is: to summarize, in a single average, statements of the type, "a units of one quantity correspond to b units of another," find the arithmetic mean of ratios in the form $\frac{a}{b}$ if the quantity b is the fixed element in the discussion. The same result may of course be secured by finding the harmonic mean of the ratios in the form $\frac{b}{a}$. In some cases, it is easy to identify which quantity is "the fixed element in the discussion"; in other cases the decision depends in part on knowledge of the customary usage and judgment as to which choice will make clear to most probable users the real meaning of the figures.

As an instance of those cases in which knowledge of customary usage and judgment are necessary, and in which there has actually been some difference of opinion, let us consider a problem in educational measurement. A class of pupils has been given 10 minutes in which to work as many as they can of a given list of problems, and the number of problems worked correctly by each pupil noted. Professor Kelley [1] thinks that the number of minutes a pupil requires to work one problem is a more "straightforward" measure than number of problems a pupil can work in a given time, and that therefore the harmonic mean of number of problems worked in 10 minutes is better than the arithmetic. It seems clear to the writer, however, and to many educational statisticians, that the *time* is the

[1] *Statistical Method*, by Truman L. Kelley, chap. III, sec. 15, p. 64.

Digitized by Google

fixed element in this discussion, and would be so even if the conditions were altered by allowing the pupils all the time necessary to work the full list of problems. As may easily be verified by reference to standard works on physics, chemistry, engineering, and calculus, rates in which time is one element are almost invariably stated as so many units of distance, temperature, electricity, or what-not per unit of time. Even in popular speech, miles-per-hour is much more common than minutes-per-mile. The writer maintains, therefore, that number of problems worked in a given time is the "straightforward" measure, and that the arithmetic mean of such measures is the correct average. In the modified version of the problem, in which results were recorded in the form, "minutes required to work complete list," we should either find the harmonic mean of these times, or convert each separately to "problems per hour" and find the arithmetic mean.

Geometric mean. There are two classes of problems in which it is necessary to use the *geometric mean* — (1) those involving interpolation in historical or quantitative functional series, in which the quantities increase in each interval by a uniform percentage of the value at the beginning of the interval, and (2) those in which it is desired to combine two or more series. As an instance of the former problem, consider what would be the best estimate of the population in 1915 of a city whose population in 1910 was 100,000 and in 1920 144,000. If the proportion of increase in each five year period was the same, then the ratio of 1915 population to 1910 population is equal to the ratio of 1920 population to the 1915 population. Or in symbols,

$\frac{P_{1915}}{P_{1910}} = \frac{P_{1920}}{P_{1915}}$, from which we easily derive $P_{1915} = \sqrt{P_{1910} \cdot P_{1920}}$

This combination of two figures is called the geometric mean; more generally for n quantities we define

$$G = \sqrt[n]{a_1 \cdot a_2 \ldots a_n}$$

This particular usage (1), however, is not so much a means of summarizing as a method which anticipates the finding of trend lines. Its logical extension is to interpolating other than at the mid-

Digitized by Google

point. If we want to find the 1914 population, for instance, we think of the population as increasing each year by a uniform percentage of its value at the beginning of the year — i.e., each year the population is multiplied by a factor R, a little greater than 1.

After 10 years the population is therefore $P_{1910}\ R^{10} = P_{1920}$ (which is essentially the formula for compound interest), and

$$P_{1914} = P_{1910}R^4$$

The determination of P_{1914} is then a simple matter of logarithmic computation.

$$\log P_{1910} + 10 \log R = \log P_{1920},$$

$$\log R = \frac{\log P_{1920} - \log P_{1910}}{10}$$

$$\log P_{1914} = \log P_{1910} + \frac{4}{10}(\log P_{1920} - \log P_{1910})$$

It should be noted that R–1 is the average rate of increase per year, on this basis. This is the correct method for computing such an average if the quantity increases in each interval by a constant percentage of its value at the beginning of that interval. It is equivalent to using the ratio chart.

If the population in 1914 were computed on the assumption that the amount of increase was a uniform amount each year rather than a uniform percentage of the amount of the beginning of each period, the equation corresponding to that just stated would be

$$P_{1914} = P_{1910} + \frac{4}{10}(P_{1920} - P_{1910})$$

That is to say, the geometric mean amounts to using logarithms instead of the quantities themselves, or graphically, to using the "ratio chart" instead of the ordinary diagram.

The second use of the geometric mean is an actual attempt to summarize divergent tendencies, as an average is supposed to do. It is related to the preceding also, however, in that the ratio concept is fundamental. The occasion for the use is this: suppose one quantity increases 100 per cent and another decreases 50 per cent. What is the best figure to take as the combined rate of increase, if the quantities are taken to be of equal importance?

Digitized by Google

The A.M. of 100 and −50 is 25, or the A.M. of the relative numbers — 200 and 50 — is 125, a 25 per cent increase. The harmonic mean of the relative numbers is given by H where $\frac{1}{H} = \frac{\frac{1}{200} + \frac{1}{50}}{2} = \frac{5}{400}$, $H = 80$, indicating a 20 per cent *decrease*. The geometric mean of the relative numbers, using the preceding definition, is $\sqrt{200 \cdot 50} = 100$, indicating no change! Suppose that in a second interval of time, each quantity returns to its original value, that is to say, the first decreases 50 per cent and the second increases 100 per cent. If we still regard the quantities as of equal importance, the average rate of increase, by the A.M. method is therefore + 25 per cent; by the H.M. method, − 20 per cent; by the G.M. method, 0. It is obvious that a method which says that two increases of 25 per cent may mean a return to the original value is defective; as is also one which says that two decreases of 20 per cent may mean a return to the original value. The G.M. is therefore for such cases as this clearly a superior means of summarizing in that it leads to sound results without the careful attention to weights which must be given if the arithmetic or harmonic means are used. This point will be discussed further in Chapter VI.

Special averages appropriate for special purposes. In addition to the more usual averages, others may be appropriate for special purposes. As an instance of such a special average, let us find the average age which should be used to summarize the ages of a group of men who are about to take out ordinary life insurance policies in a certain company. The appropriate average age to use for this purpose is clearly such that if all the men were of that age, the premiums would total exactly as much as they do with the actual distribution. Let us see how this principle works out in a numerical case.

Let us suppose the group consists of one man of 20, two of 25, three of 30, three of 35, three of 40, two of 45, and one of 50, and that the premiums for each age are as follows:

20	13.48	32	18.26	36	20.84	40	24.21
25	15.10	33	18.84	37	21.60	45	29.77
30	17.19	34	19.45	38	22.42	50	37.79
31	17.72	35	20.11	39	23.29		

Digitized by Google

It will be found that the average premium for this group of men at the rates quoted is 21.70, which very nearly corresponds to the premium for age 37. We therefore say that the *effective average* age for insurance purposes is 37. It is evident on inspection that the average age for most purposes is 35. It should also be noted that the average age for the purpose of other types of insurance, or for insurance in another company, may differ from that found. The question naturally arises, why use an average of this sort, which has to be found by the same arithmetic which one would use in doing the problems for which it can be expected to be of value? The answer is, that the average merely assists one to form a comprehensive mental picture of the group, and thus may make it easier to decide questions which may arise.

MISCELLANEOUS EXAMPLES

1. The following table indicates the number of units purchased at various prices. How many units would have to be purchased at 30 cents apiece so that the average price would be 29 cents?

Price (cents)	15	18	21	24	27	30
No. of units	2	9	10	35	50	?

2. The following table gives for four students the grade on work done during the semester and the grade on the final examination. Find the general average grade for the semester, if the former is to be weighted 3 and the latter 1. Do by the method of "averaging in."

Student	Semester Grade	Final Examination	General Average
A	78	54	
B	61	51	
C	66	76	
D	77	67	

3. Find the average grade, if the average for four tests is given in column I and for a fifth test in column II.

Student	Column I	Column II
A	52	72
B	60	50
C	80	65
D	93	78

4. Estimate the population of the United States in 1917 by the geometric mean method. (Compare Chapter II, page 31.)

5. A man bought car checks at two different times. The first time he received 10 car checks for each 50 cents, and the second time 9 for each 50 cents.

Digitized by Google

Find the average price, first on the assumption that he purchased 90 checks on each occasion, and second on the assumption that he purchased $5.00 worth of checks on each occasion. Explain how the harmonic mean could be used in this problem. Ans. (*a*) 5.28¢; (*b*) 5.26¢.

6. The following table shows the number of associate professors at certain American colleges and universities, whose salaries fell in the classes indicated. Note the modal salary class, and find the median and arithmetic mean salaries. In your judgment, which average is most typical?

Salary Class (by mid-point)	Number	Salary Class (by mid-point)	Number
250	1	2250	168
500	3	2500	174
750	3	2750	129
1000	4	3000	153
1250	15	3250	74
1500	57	3500	91
1750	88	3750	17
2000	186	4000	15
		4500	1

7. The prices of two articles in certain years were as follows:

	1914	1918	1922
A	.32	.40	.44
B	1.20	.96	1.08

Find for 1918 and for 1922 the geometric mean of the ratios of increase over 1914.

8. The A.M., median, and modal class should be found for a considerable number of the frequency distributions given in the examples at the ends of Chapter IV and Chapter VII.

Answers: VII–1 Mean 3118, Median 2971.
VII–2 Mean 80.14 Median 75.4.

Comparison of the averages. The main aims of the preceding portion of this chapter have been to define the five principal averages, to explain how to compute each, and to discuss briefly in each case one or more types of problem for which that average is the most significant typical value. We have not yet, however, discussed the task of selecting one or more of these averages from the point of view of the practical statistician who desires to discover and explain to others the true meaning of certain data. In many such cases the selection is controlled by the principles discussed above in connection with particular averages. We have seen, for instance, that the arithmetic mean is closely related to the total of the magnitudes. If the total of the magnitudes is a significant figure, the arith-

Digitized by Google

metic mean is probably the most valuable average, and the weights should be those which make the total most significant. If it is the total of the *reciprocals* of the magnitudes which is significant, the harmonic mean is preferable. If a few unusually large or small items are included, as often happens in wage or salary distributions, the median may be desirable as a substitute for or in addition to the arithmetic mean.

There are many cases, however, in which such general rules do not apply. In such cases the selection of an average may be facilitated by a knowledge of certain relationships which we shall discuss in the next sections.

Relative size of A.M., G.M., and H.M. Let A, G, and H be respectively the arithmetic, geometric, and harmonic means of a and b. Then

$$A = \frac{a+b}{2},\ G = \sqrt{ab},\ \frac{1}{H} = \frac{\frac{1}{a}+\frac{1}{b}}{2},\ H = \frac{2ab}{a+b}$$

If a and b are any two different positive numbers, $(\sqrt{a} - \sqrt{b})^2$ must be positive;

That is, $(\sqrt{a} - \sqrt{b})^2 > 0$ (I)

Expanding, we have $a - 2\sqrt{ab} + b > 0$ (II)

Transposing, we have $a + b > 2\sqrt{ab}$ (III)

Dividing by 2, we have $\frac{a+b}{2} > \sqrt{ab}$; i.e., $A > G$.

Multiplying both sides of (III) by $\sqrt{ab}$ and dividing by $a+b$, we have:

$$\sqrt{ab} > \frac{2ab}{a+b};\ \text{i.e.,}\ G > H.$$

Or, in words, we have shown that the arithmetic mean of two different magnitudes is always greater than their geometric mean, which in turn is greater than their harmonic mean.

Suppose now we have the arithmetic, geometric, and harmonic means of n quantities, not merely of 2; to prove that again $A > G > H$.

$$A = \frac{a_1 + a_2 + \ldots a_n}{n},\ G = \sqrt[n]{a_1 a_2 \ldots a_n},\ \frac{n}{H} = \frac{1}{a_1} + \frac{1}{a_2} + \ldots \frac{1}{a_n}.$$

Digitized by Google

If all the a's are equal, $A = G = H$. If not, replace in both A and G the largest a (a_i) and the smallest a (a_k) by $\frac{a_i + a_k}{2}$. A is unaltered by this change; G is altered to G_1 by replacing $a_i a_k$ by $\frac{(a_i + a_k)^2}{4}$. Since as proved above, $\sqrt{a_i a_k} < \frac{a_i + a_k}{2}$, we have after squaring, $a_i a_k < \frac{(a_i + a_k)^2}{4}$. Therefore $G < G_1$. We repeat this process — i.e., replace the largest and smallest a's each by their arithmetic mean, getting a new $G(G_2)$, which is $> G_1$. We have then $G < G_1 < G_2 \ldots$, carrying on this process. The a's in each successive G are more nearly equal; the limiting G.M. (say Γ) approached by this process is one in which all the a's are equal to each other and to A. Then $\Gamma = \sqrt[n]{A.A \ldots A \text{ to } n \text{ factors}} = A$. We have therefore $G < A$.

Similarly to compare H and G, if we replace each of the terms $\frac{1}{a_i}$ and $\frac{1}{a_k}$, corresponding to the largest and the smallest a's, by $\frac{\frac{1}{a_i} + \frac{1}{a_k}}{2}$, the value of H is unchanged, but $a_i a_k$ in G becomes $\frac{4a_i^2 a_k^2}{(a_i + a_k)^2}$. But it was proved above that $\sqrt{a_i a_k} > \frac{2a_i a_k}{a_i + a_k}$, and hence $a_i a_k > \frac{4a_i^2 a_k^2}{(a_i + a_k)^2}$. The new G (G_1') is therefore less than the original. We therefore have a series $G > G_1' > G_2' > G_3' \ldots$ etc., in which the elements of the successive G's are more nearly equal. The limiting G (say Γ') approached by this process is one in which all the a's are equal to each other and to H.

$$\text{Then } \Gamma' = \sqrt[n]{H \ldots H \text{ to } n \text{ factors}} = \text{H.}$$

We have therefore $G > H$.

Effect of extreme items on the different averages. The five averages are affected in characteristically different ways by items which are markedly larger or smaller than the rest of the group.

Digitized by Google

The mode is not changed at all by replacing an item near the center of the group, but not actually at the mode, by a very small or a very large item. The median is changed only if the new item is on the other side of the original median than the item it replaces. The arithmetic, geometric, and harmonic means are appreciably changed by such replacement, the exact extent of the change depending of course on the number of items and on the magnitude of the difference between the original and replacing items, and being different ordinarily for each of these three averages. It should be noted that neither the geometric nor the harmonic mean is significant if zero is one of the items, and that these averages are not typical of the group of items if one or more of the items is very small relatively to the rest.

By a consideration of the appropriate algebraic formulas, the following general principles may be established:

(1) If an item larger than the harmonic mean is replaced by a still larger item, the arithmetic mean will be increased more than the harmonic.

(2) If an item less than the harmonic mean is replaced by a still smaller item, the harmonic mean will be decreased more than the arithmetic.

These principles may be illustrated, and the characteristic "middle of the road" behavior of the geometric mean emphasized, by noting these means for the following three sets of items:

Set I. 4, 5, 5, 6, 6, 8, 9, 10, 10, 10, 12, and 15.
Set II. Set I with 18 substituted for 9.
Set III. Set I with 1 substituted for 6.

The results are as follows:

	A.M.	G.M.	H.M.	Change as compared with Set I	
				A.M.	H.M.
Set I.	8.33	7.75	7.19		
Set II.	9.08	8.21	7.44	+.75	+.25
Set III.	7.92	6.67	4.80	−.41	−2.39

These principles may be stated less precisely, but in a form which is helpful for estimating what differences may be expected in the averages because of differences in the groups:

Digitized by Google

The arithmetic mean is most affected by a few unusually large items, and the harmonic most by a few unusually small items. The effect on the geometric mean is intermediate between the effect on the A.M. and the effect on the H.M.

As an example of the relative effect on mean, median, and mode of replacing some items by others, let us consider again the frequency distribution:

Length of ear (inches)	3	4	5	6	7	8	9	10	11	12
Number of ears	1	1	3	14	25	62	119	80	19	2

As it stands the modal class-interval of this series is 8.5 to 9.5 inches, or briefly, 9 inches, the median 8.98, and the A.M. 8.84. If one ear of eight inches is replaced by one of 40 inches, the mode is unchanged, the median becomes 8.99, and the A.M. 8.94. Again, if 20 9-inch ears should be replaced by 20 10-inch ears, the modal class-interval becomes 10 inches (strictly, 9.5–10.5 inches), the median becomes 9.08, and the A.M. 8.90. As these instances indicate, it is not possible, without detailed examination of the circumstances, to say whether the median or the A.M. will be most affected by a shifting of items.

Ease of computation and manipulation. Of all the averages, the median is the one whose exact value is most easily found unless the distribution includes over 100 or 150 items. The mode is easier in the case of discrete distributions but in the case of continuous distributions, only the modal group can be located exactly by methods so far considered, the exact value being indeterminate. The arithmetic mean can be computed, by the deviation method, or by direct summation with a computing machine, with little real difficulty, but the geometric and harmonic means are more troublesome, enough so that they are not worth finding except when it is clear that they give valuable information.

The arithmetic mean is much superior to the mode and the median as a basis for algebraic discussion. As shown by some of the examples considered in this chapter, certain minor changes in the position of particular items in the series will at times cause abrupt changes in the mode or the median, and such changes in-

Digitized by Google

terfere seriously both with the use of these averages and with the theoretical discussion of their properties.

Which should be used? If in doubt as to which average it is best to use in a given problem, think, which best summarizes the frequency distribution? Which is the value that would least misrepresent the facts if the reader thinks of all the values concentrated at the point instead of spread out as they really are?

It is better, of course, to use two or three averages rather than one, especially when writing for an audience familiar with statistical methods. By so doing we describe the frequency distribution to some extent. In certain cases other figures from the full table may be even more valuable than the averages. If we are summarizing, for instance, the amount of training of a group of teachers, the figure for percentage normal trained is rather better than that for average number of years of training. In general, in using averages, bear in mind that the aim of all statistical discussion is to make the facts more available for deciding questions that have arisen or are likely to arise.

Summary of chapter. An *average* is a typical number selected to represent the magnitude of a measured quantity in a group of individuals. It summarizes a frequency distribution as much as is possible in a single figure. It should not be used, however, for detailed discussion which really requires the frequency distribution, stated in full.

The *arithmetic mean*, usually called the average, is the sum of the measures of a certain characteristic for a group of individuals, divided by the number of individuals. It may be computed by the formulas

$$\bar{x} = \frac{\Sigma x}{n}, \quad \bar{x} = \frac{\Sigma fx}{n}, \quad \bar{x} = \Sigma wx.$$

In some cases it is convenient to compute the average by assuming an approximate value, finding the average of the differences from this approximate value, and applying the result as a correction to the assumed average.

Digitized by Google

A second average, based on a different set of weights, will be larger than the first if the second set of weights has on the whole meant a greater stressing of the larger measures and a smaller stressing of the smaller measures.

The *median* is the magnitude of the measured characteristic of the middle item, when all the items are arranged in the order of size of that characteristic. The median of a frequency distribution is located, in the absence of more detailed information, by assuming that the individual magnitudes are uniformly distributed over the median interval. Such exact location of the median is not legitimate if all the items of the median interval are, by the conditions of measurement, concentrated at a limited number of points.

The median should be used if comparison with the whole frequency distribution shows that it is a typical number. In some cases it should be used with the arithmetic mean to give a compact summary of the distribution.

The *mode* is the value of the measured characteristic which occurs most frequently. It corresponds rather more closely than the arithmetic mean or the median to ordinary thoughts or remarks about what "usually" happens, and is useful for popular presentation of statistical results. Attempts at exact location of the modal point, rather than the modal interval, are not recommended.

The *harmonic mean* is the reciprocal of the arithmetic mean of the reciprocals of the individual measures. It should be used in certain problems in which the quantities which it is convenient to average are the reciprocals of the quantities of which the arithmetic average is really required by the conditions of the problem.

The *geometric mean* is the n'th root of the product of the n measures. It is used to interpolate between two points in a historical series in which the growth in an interval is a constant percentage of the value at the beginning of the interval, and also to combine two or more historical series so as to secure certain properties in the resulting series.

For special purposes, averages must be computed in other ways than those enumerated above.

The arithmetic mean is always larger than the geometric mean, which is in turn larger than the harmonic mean.

Digitized by Google

The arithmetic mean is much affected by a few extremely large items, and the harmonic mean by a few extremely small items. The median is likely to change abruptly for small shifts in the items, in the case of a discrete distribution.

The easiest average to compute exactly is the median, with the arithmetic mean not far behind. The arithmetic mean can be treated algebraically with the greatest ease and completeness.

If in doubt as to which average to use, choose the one which would least misrepresent the facts if all the values were thought to be concentrated at that point rather than spread out as they really are. Try to visualize the problems in which the average will be used, and assist those who will try to solve these problems by giving them the average (or averages) which will meet their needs.

Digitized by Google

CHAPTER VI

INDEX NUMBERS

THE following problem typifies those which often confront a business man or economist:

A manufacturer of brass rod notes that the price of copper is 20 per cent higher this year than last, and the price of zinc is 5 per cent higher. He wants to know what is the *average* increase in the price of his raw materials.

A little reflection or trial computation will show that no definite answer can be given to his question without certain additional information. As is often the case, the problem really includes determining what information is needed, as well as deciding on the method of computation to be used, securing the information, and doing the computation.

First method of solution of typical problem. One method of solving the problem, which is clearly correct, though perhaps not the simplest method, is as follows:

(1) Find what the materials used last year cost at last year's prices.
(2) Find what the same quantities would cost at this year's prices.
(3) From the latter result subtract the total found in (1), thus getting the increase in cost due to price change.
(4) Divide this increase in cost by last year's cost, as found in (1), and multiply by 100, thus getting per cent increase in costs.

For the sake of making it clear just what is to be done and what additional data are needed, and of introducing the algebraic notation which experience has shown to be valuable in such problems, we shall carry out the details of the computation in full.

Let us suppose that the data are as follows:

Digitized by Google

Quantity of copper purchased ($= q_1$)	100,000 pounds
Quantity of zinc purchased ($= q_2$)	50,000 pounds
Price of copper last year ($= p_1$)	15 cents per pound
Price of zinc last year ($= p_2$)	10 cents per pound
Price of copper this year ($= p'_1$)	18 cents per pound
Price of zinc this year ($= p'_2$)	10.5 cents per pound

We deduce at once:

Cost of copper at last year's prices (p_1q_1)	\$15,000
Cost of zinc at last year's prices (p_2q_2)	\$5,000
Total cost of raw materials at last year's prices ($p_1q_1 + p_2q_2$) or Σpq	\$20,000
Cost of same amount of copper at this year's prices (p'_1q_1)	\$18,000
Cost of same amount of zinc at this year's prices (p'_2q_2)	\$5,250
Total cost of raw materials at this year's prices ($\Sigma p'q$ or $p'_1q_1 + p'_2q_2$)	\$23,250
Increase in cost ($\Sigma p'q - \Sigma pq$)	\$3,250
Per cent increase $\left(\frac{(\Sigma p'q - \Sigma pq)}{\Sigma pq} \times 100\right)$	16.25

Ratio of this year's cost to last year's or the *index number* $\left(\frac{\Sigma p'q}{\Sigma p\,q}\right)$ i.e., 23,350/20,000 = 116.25 per cent

Second method of solution of typical problem. We may note, using algebraic symbols, that the ratio of increase in copper price was

$$\frac{p'_1 - p_1}{p_1} = \frac{p'_1}{p_1} - 1 \ (= .20)$$

and in price of zinc,

$$\frac{p'_2 - p_2}{p_2} = \frac{p'_2}{p_2} - 1 \ (= .05)$$

Let us try to find the average rate of increase by weighting each of these ratios of increase by the corresponding proportion of total

Digitized by Google

costs at last year's prices. From the details shown above, we find that three fourths of the cost was for copper and one fourth for zinc. The average ratio of increase is therefore

$$\tfrac{3}{4} \times .20 + \tfrac{1}{4} \times .05 = .1625, \text{ or } 16.25 \text{ per cent.}$$

Algebraically, this average ratio of increase is:

$$\frac{p_1 q_1}{p_1 q_1 + p_2 q_2}\left(\frac{p_1'}{p_1} - 1\right) + \frac{p_2 q_2}{p_1 q_1 + p_2 q_2}\left(\frac{p_2'}{p_2} - 1\right).$$

This may be reduced readily to the form

$$\frac{\Sigma p'q - \Sigma pq}{\Sigma pq}$$

and the exact correspondence with the result of the first method verifies the second method.

Definition of index number. An *index number*, then, is an average which combines two or more historical series, such as the two price series in the preceding example. A price index number may be regarded either as:

(1) The ratio of the cost of certain quantities of certain articles at one date to their cost at another date; or
(2) The average ratio of prices at one date to prices at another date.

Index numbers as actually used are not limited to a comparison of two dates, but the ratios of costs (or prices) for a succession of dates, to costs (or prices) for a fixed date, are computed and referred to as an index number with fixed year as base. In current discussions, it will be found that the phrase "index number" is used to mean (1) a particular ratio, or (2) a series of similar ratios, or (3) the measure of price variation which such a series constitutes. In the latter sense, an index number is considered to be defined if we know (1) the articles whose prices were used, (2) the method of selecting the prices at each date, (3) the weights, (4) the base month, year, or other period, and (5) the method of combination or the algebraic formula used.

Digitized by Google

Price index numbers have sometimes been constructed by comparing prices at one *place* with prices at another *place* rather than prices at one date with prices at another date. We may thus compare, for instance, the food cost of living in two different countries. This use is much less common, however, because it is more difficult to secure comparable prices for different places at the same time than for different times in the same place.

Index numbers have been used to measure changes in production as well as price. A production index number may be regarded either as:

(1) The ratio of the aggregate value, at some fixed set of prices, of the quantities of certain articles produced in one period to the value at the same prices of the quantities produced in another period.
(2) The average ratio of quantities produced in one period to quantities produced in another period.

Noteworthy points in the preceding example. The example of a price index number discussed above is unusually simple in that only two items and two dates were considered, and the weights were the same for both dates. A thorough understanding of the ideas involved, however, and of the two methods of computation, will prepare the reader for the complications in detail which arise in connection with index numbers actually in use. Attention is called to the following points:

(1) Index numbers are usually quoted in the form of relative numbers with the average or total for the first or other early year taken as base. In fact some writers use the term *index number* for what we have called *relative number*, and speak of the concept we are discussing in this chapter as an *average index number*, when they wish to make the distinction. The terminology we have used seems preferable, however, as the scheme of using a single series of numbers as an index or indicator of more complicated data, does not require that the series be reduced to relative numbers. Some of the oldest index numbers are totals or averages primarily, and yet are

Digitized by Google

called index numbers by their originating source. Bradstreet's index number, for instance, is just the average price per pound of a selected list of raw materials, foodstuffs, and other items. The index number of British prices prepared by the London Economist, also, is stated primarily as a total.

(2) If an index number is computed as the ratio of total cost at one date to total cost at another, total cost is of course computed by weighting prices by quantities used, purchased, or sold. In the above example it was assumed that these quantities were the same at both dates; if there has been any change, a further difficulty is presented which will be considered later.

(3) If on the other hand an index number is computed as an average of price-ratios, as in the second method above, the appropriate weight for each price-ratio is the total cost of the quantity of the corresponding article purchased at the price of the base date; i.e., the price which is the denominator of the price-ratio. As was shown algebraically above for two items (and the same proof could be extended to any number of items), such weights lead to exactly the same final results as are obtained by the first method.

(4) Instead of the actual quantities in (2) or actual costs of quantities purchased in (3), any set of numbers proportional to them may be used. This is obvious on inspection of the algebraic formulas, if we replace each q by tq and then note that on simplification the original formulas are reproduced.

In particular, the arithmetic is simplified if we select the factor t as equal to the reciprocal of the sum of the quantities (or costs), thus making the sum of the weights equal to unity.

EXAMPLES

Do each of the following examples both by Method I and by Method II, if possible with the data given. Bear in mind the principles in regard to weights summarized in points (2) and (3) above.

1. A man of simple tastes lives on bread, milk, eggs, and bananas. Each week he consumes 3 loaves of bread, 7 quarts of milk, 2 dozen bananas, and one dozen eggs. If the prices were as follows on the first day of January, 1924 and

Digitized by Google

1925, respectively, compute the index number of the cost of his food, January 1, 1925, taking January 1, 1924, as base.

	Bread (loaf)	Milk (qt.)	Eggs (doz.)	Bananas (doz.)
1/1/1924	12 cents	15 cents	50 cents	20 cents
1/1/1925	15 cents	12 cents	60 cents	30 cents

2. A certain manufacturer requires for a unit of product the following raw materials: 1 ton of steel billets, 20 pounds of tin, and 50 pounds of copper. If the prices on certain dates were as indicated below, find the index number of his cost of raw materials, first with 11/1/1913 as base, and then with 11/1/1920 as base.

	Copper (lb.)	Steel Billets (ton)	Tin (lb.)
11/1/1913	16.5¢	22.00	40¢
11/1/1920	18.5¢	55.00	43.5¢
11/1/1924	13.5¢	35.50	53¢

3. A tire manufacturer finds that the prices he has been paying for rubber are 100 per cent higher than they were a year ago, while cotton webbing is 10 per cent higher. If his purchases of cotton webbing last year had a value one half that of his purchases of rubber, find the index number of the prices of his raw materials, last year as base. It is assumed that these two items were the only raw materials purchased.

4. A man built a greenhouse in 1913 at a cost of $100 for window-glass and $200 for lumber. If lumber had doubled in price and window-glass was 52 per cent higher, in 1924, what was the average percentage of increase of cost for these raw materials?

5. Compute an index number of agricultural production in the United States from the following data:

Annual Production

	Corn (bu.)	Wheat (bu.)	Cotton (bales)	Hay (tons)
1913	2,447,000,000	763,000,000	14,160,000	80,500,000
1918	2,503,000,000	921,000,000	12,040,000	91,100,000
1923	3,054,000,000	786,000,000	10,140,000	106,600,000

Approximate average prices, 1920–1923 inclusive:

Corn, per bushel 62 cents
Wheat, per bushel, $1.07
Cotton, per bale, $106
Hay, per ton, $13.20

First complication: variation in quantities. All the problems considered so far in this chapter, including the five examples above, involve the assumption that the weights either are constant or may be taken as constant for the purpose in hand. In example (5), for instance, the prices, which are used as weights in the computation, are the average prices for a recent period. In many problems which actually arise, however, the natural weights are different for the different dates.

Digitized by Google

As an example which involves this difficulty, let us attempt to construct an index number of wages in a certain factory in which the relative number of workers of different types changes between the dates considered. Let us suppose that on January 1, 1924, there were 100 laborers, with earnings of $25 a week, 400 semi-skilled workers with earnings of $30 a week, and 500 skilled workers with earnings of $35 a week. On January 1, 1925, there were 50 laborers, earning $27 a week, 100 semi-skilled workers earning $29 a week, and 650 skilled workers earning $33 a week. What is the average percentage change in wages?

There are several methods of computation which seem at first thought to be of possible significance. Let us compute the rate of change according to what seem to be the most promising of these methods, and state the algebraic formula for each method. In these formulas, let us denote by q_1, q_2, q_3, the number of employees of the three grades, January 1, 1924, by p_1, p_2, p_3, the wage rates, January, 1924, and by q_1', q_2', q_3', and p_1', p_2', p_3', the number of employees of each class and the wage rates, January 1, 1925. We have, then:

METHOD 1. *Ratio of pay rolls, i.e.,* $\frac{\Sigma q'p'}{\Sigma qp}$

The weekly pay roll in January, 1924, was:

$$100 \times 25 + 400 \times 30 + 500 \times 35 = 32{,}000$$

and in January, 1925:

$$50 \times 27 + 100 \times 29 + 650 \times 33 = 25{,}700$$

The pay roll in 1925 was therefore only 80.3 per cent of what it was in 1924.

METHOD 2. *Ratio of average earnings per employee, i.e.,* $\frac{\frac{\Sigma q'p'}{\Sigma q'}}{\frac{\Sigma qp}{\Sigma q}}$

Dividing each pay roll above by the number of employees at that date, we find that the average weekly earnings per employee were $32.00 in 1924 and $32.13 in 1925. The per capita average wage in 1925 was therefore 100.4 per cent of what it was in 1924.

METHOD 3. *Ratio to actual* 1924 *pay roll, of pay roll at* 1925 *rates to* 1924 *employees, i.e.,*

$$\frac{\Sigma qp'}{\Sigma qp}$$

Digitized by Google

If the number of employees of each grade had been the same in 1925 as in 1924, the weekly pay roll, at the 1925 rates, would have been:

$$100 \times 27 + 400 \times 29 + 500 \times 33 = 30{,}800$$

Comparing this with the actual 1924 pay roll of $32,000, we see that, because of wage changes alone, the 1925 pay roll would be 96.2 per cent of the 1924 pay roll, or in other words, that the average wage rate has declined 3.8 per cent, from this point of view.

METHOD 4. *Ratio of actual* 1925 *pay roll to pay roll at* 1924 *rates to* 1925 *employees, i.e.,*

$$\frac{\Sigma q'p'}{\Sigma q'p}$$

If the number of employees of each grade in 1924 had been what it became in 1925, the weekly pay roll, at the 1924 rates, would have been:

$$50 \times 25 + 100 \times 30 + 650 \times 35 = 27{,}000$$

Comparing this with the actual 1925 pay roll, $25,700, we see that, if the weights had been at both dates as in 1925, the 1925 pay roll would have been 95.2 per cent of the 1924 pay roll.

METHOD 5. *Ratio of* 1925 *pay roll to* 1924 *pay roll, if number of employees of each grade had been the average of* 1924 *and* 1925 *numbers, i.e.,*

$$\frac{\Sigma \frac{q+q'}{2} \times p'}{\Sigma \frac{q+q'}{2} \times p}$$

As a compromise between Method 3 and Method 4, let us take the number of employees of each grade as the average of the number in 1924 and the number in 1925, namely, 75 laborers, 250 semi-skilled workers, and 575 skilled workers. Then at the 1924 rates, the pay roll would have been:

$$75 \times 25 + 250 \times 30 + 575 \times 35 = 29{,}500$$

At the 1925 rates, the pay roll would have been:

$$75 \times 27 + 250 \times 29 + 575 \times 33 = 28{,}250$$

This hypothetical 1925 pay roll is therefore 95.8 per cent of the corresponding 1924 pay roll, or in other words, the average rate has declined 4.2 per cent, from this point of view.

METHOD 6. *Weighted average ratio of change in wage rates, i.e.,*

$$\Sigma \left[\frac{qp}{\Sigma qp} \left(\frac{p'-p}{p} \right) \right]$$

All of the preceding methods are variations of the first method used in the first example. It is also possible to use variations of the second method. For instance:

We note that the rate of increase of laborers' wages has been 8 per cent, that

Digitized by Google

the rate of decrease in unskilled workers' wages has been 3.3 per cent, and that the rate of decrease in skilled workers' wages has been 5.7 per cent. To combine these rates, we should use as weights proportions of pay roll based on the 1924 wage rates. In making up these weights, however, while we must use 1924 wage rates, inasmuch as we have based the percentages of change on 1924, yet we may use either 1924 or 1925 numbers of employees, or may use some compromise. For the sake of illustration, let us use the 1924 number of employees. The combined rate of change is:

$$\frac{2,500}{32,000} \times 8 + \frac{12,000}{32,000}(-3.3) + \frac{17,500}{32,000}(-5.7) = -3.8 \text{ per cent}$$

which agrees with the result of Method 3 above. As the student may verify, use of the 1925 number of employees leads to the same result as Method 4, and use of the average number of employees leads to the same result as Method 5.

Discussion of various methods. The results obtained by the first five methods differ enough so that we cannot possibly regard them as substantially equivalent. How are we to decide which of them will best serve some specific purpose? Can any one of them be regarded as a satisfactory "general purpose" index number of change in wage rates?

The best way to get at this question is to consider carefully several problems in which an index number might be of value, and to select in each case that one of the five which throws the most light on that problem. One problem, for instance, is to measure changes in the purchasing power of employees of the given factory. For this purpose, changes in number of employees are as important as changes in wage rates, and the comparison of total pay rolls provided by Method 1 is more useful than the results of the other methods.

A second problem is to measure the average change in wage rate between the two dates. The most natural measure of this change is derived by Method 3, comparing what the workmen actually received per week in 1924 with what the same number of men of each grade would have received at the 1925 rates. To be sure, it is almost equally as natural to make this comparison on the basis of the number of men of each grade on the roll in 1925, as done in Method 4. As long as only two dates are being compared, it is convenient to compromise, either by using average weights, as done in Method 5, or by taking the geometric mean of the results of Methods

Digitized by Google

3 and 4. The last-named scheme is recommended by Professor Irving Fisher [1] as giving the "Ideal Index Number," the formula being:

$$I = \sqrt{\frac{\Sigma qp'}{\Sigma qp} \times \frac{\Sigma q'p'}{\Sigma q'p}}$$

If there were reason for believing that the distinctions between the different grades of workers were not the same at the two dates, it would not be worth while to make the special computations required for Methods 3, 4, or 5, as the simpler computation of Method 2 would be just as satisfactory. To be sure, it ignores the possibility that there was a variation in the composition of the group of employees; but it is better to ignore such a distinction than to attempt to carry through refinements not justified by the data at hand. As a matter of fact, most business men would use Method 2 almost without question, largely because reliable data on the composition of the working force are seldom available. They would recognize that comparisons between two factories or two industries on the basis of average wage per employee might be invalidated by differences in the composition of the working force, but would be inclined to assume without further examination that such differences would be inconsiderable in comparing wages in the same factory at two dates. This assumption would sometimes be in error, but would be right often enough to justify the use of Method 2 for approximate purposes.

A third problem in which these wage indices might be of value is that of reducing the costs of a later year to the cost level of an earlier year. If, for the sake of the present illustration, it may be assumed that labor costs vary directly with wages, then it is obvious that the relation between actual labor costs in 1925 and what they would have been in 1924 for the same workers, is given by the result of Method 4. We can therefore reduce amount spent for labor in 1925 to the 1924 wage level by dividing by this index. Or from the other point of view, if we want an index of wage increase which could be used to reduce 1925 costs to the 1924 level, we should use Method 4.

[1] *The Making of Index Numbers.*

Digitized by Google

For the sake of variety in subject-matter, we have used number of employees and wage rates in discussing how to deal with variable weights. Substantially the same considerations apply if quantities are substituted for number of employees and prices for wage rates, except that Methods 1 and 2 have to do with total and average costs of quantities exchanged and would not be used by any one as price indices.

How important is this "first complication: variation in quantities"? In most practical problems, there will actually be some differences in the relative quantities at the two dates. The only safe way to be sure that such differences are not significant is to compute the index by both Methods 3 and 4 and note the difference. If the difference is considerable, choice between them or Method 5 should be made according to the purpose in hand or the point of view of the prospective user, and the statement of the result should make clear what method has been used. If the weights do not change much and the price changes are fairly uniform for the various items, the results of all three methods will be nearly the same, and any one of them can be used when an index of price change is needed.

EXAMPLE

Find the index number of the food cost of living in the United States as compared with England, given the following data:

Families of the same size and relative social standing are found to have the following weekly dietaries in the two countries:

	England		United States	
	Amount	Price	Amount	Price
Bread	32 lbs.	4 cts. a lb.	12 lbs.	6 cts. a lb.
Eggs	2 doz.	30 cts. a doz.	4 doz.	35 cts. a doz.
Steak	1 lb.	50 cts. a lb.	5 lbs.	45 cts. a lb.
Beef — cheaper cuts	5 lbs.	20 cts. a lb.	2 lbs.	30 cts. a lb.
Fresh vegetables	10 lbs.	10 cts. a lb.	20 lbs.	8 cts. a lb.
Potatoes	20 lbs.	3 cts. a lb.	10 lbs.	4 cts. a lb.

Figures as stated above are for purposes of illustration, merely. They are not of course truly representative in any case, because of the small number of items considered.

Second complication: More than two dates, as well as variable weights. It was noted in connection with the first problem, in

Digitized by Google

which the quantities used as weights were the same at both dates, that the same methods were applicable to compare prices for a succession of dates with prices at a fixed date. The problem is more complicated, however, when the quantities appropriate for use as weights at three or more dates are all different. It is necessary to decide whether to use the weights of the first, second, third, or other year, or some one of the possible sets of average weights. If the purpose of the desired index number is to show what the same market basket of food (or other group of articles) would have cost at each of a series of dates, it is usually best to base the weights on the average experience of several recent years. The average for the entire period might be considered significant for some purposes, but ordinarily users of the index number will be interested most in results based on quantities as they have been in recent times.

When the question of weights is being decided, a preference is often expressed for an index number showing how the cost of the market basket actually purchased in each year compares with what the same thing would have cost in one of the years, selected as "base year." As will be explained later, this index number has its uses, but should not be allowed to usurp the place of the index number with constant weights described in the preceding paragraph. The latter may be described as a *price level* index number, and indicates the effect of price changes unobscured by the effect of changes in weights.

In order to bring out what the preceding discussion means, let us suppose that the prices paid by a certain dealer for cheese, butter, and sugar were as indicated below, and also that the quantities purchased per week were as indicated in the right hand columns.

	Prices per Pound			Quantities Purchased per Week		
	1913	1920	1925	1913	1920	1925
Cheese	15.5¢	28.5¢	27¢	100 lbs.	120 lbs.	100 lbs.
Butter	32.	61.5	50.5	100	100	150
Sugar	4.35	13.	5.	500	300	600
Algebraic symbol	p	p'	p''	q	q'	q''

If we use the quantities purchased in 1925 as weights, that is to say, use the formula $I = \frac{\Sigma q'' p'}{\Sigma q'' p} \times 100$ for 1920 and $I = \frac{\Sigma'' p''}{\Sigma q'' p} \times 100$

Digitized by Google

for 1925, we find that the price index number for 1920 is 221.8 and the index number for 1925 is 148.2. If we use the quantities purchased in 1920 as weights, that is to say, use the formula $I = \frac{\Sigma q'p'}{\Sigma q'p} \times 100$ for 1920 and $I = \frac{\Sigma q'p''}{\Sigma q'p} \times 100$ for 1925, we find that the price index number for 1920 is 211.6 and for 1925 is 153.8. The former set of index numbers shows how costs would have varied if the amounts purchased had always been the same as in 1925. It may be objected that actually the high price of sugar in 1920 reduced purchases of that commodity, and that therefore the price increase effective in 1920 was not really as great as indicated by the figure 221.8. This objection, however, does not make it any the less true that 221.8 gives a correct picture of the price situation, although it does point out that price changes alone do not tell the whole story of costs.

It must be admitted, of course, that if high price leads to the replacement of one article by another, no index number based on a constant list of items with constant weights is satisfactory. This point will be discussed more at length in a later section of this chapter.

As mentioned above, a method sometimes suggested is to compare the cost of actual purchases in a given year with what the same thing would have cost in the base year. The formulas for this method, as applied to the example discussed above, are $I = \frac{\Sigma q'p'}{\Sigma q'p} \times 100$ for 1920 and $I = \frac{\Sigma q''p''}{\Sigma q''p} \times 100$ for 1925, and the resulting price indices are 211.6 for 1920 and 148.2 for 1925, with 1913 as base. The comparison of average prices at consecutive dates by this method is obscured by the combination of changes in weights with changes in prices. The method, therefore, can not be trusted to measure reliably the changes in price level.

Use of price index numbers to eliminate effect of price change from sales. In the preceding section, it has been noted that index numbers are used most frequently to show the *price level* at a succession of dates, and that for that purpose the same set of quantities

Digitized by Google

should be used as weights throughout. A different problem in which price index numbers are sometimes used is to determine what sales (or purchases, or imports, or other similar aggregates) would have been in each of a series of years, if prices had remained stationary.

When substantially this problem was considered in an earlier section under the head of index of production, the method of solution suggested was to multiply the quantities for each year by constant prices per unit, add the products, and divide each total by the total for the base year. It sometimes happens, however, that this method is impossible or undesirable, either because the records of quantities in earlier years are incomplete, or because so many articles are involved that the computations would be prohibitively laborious. In such cases we may be able to obtain volume of sales in each year at the prices of the base year, by the following procedure:

(1) Divide total sales into classifications within each of which the trend of prices has been substantially uniform.
(2) Determine for each classification an index number of prices based on typical items. As the price trend within the class is substantially uniform, the question of the weights to be used with each item will not be very important, and can be decided on the basis of recent experience, for which the records are probably more complete.
(3) Divide the total for each class in each year by the appropriate price index, regarded as a ratio rather than a percentage.
(4) Add the class-quotients for each year, thus obtaining the volume of sales for that year in terms of the prices of the base year.
(5) If an index number for "total sales" on this basis is desired, divide actual sales for each year by the volume of sales for that year at base year prices, as found in (4). This index number is affected by changes in relative importance of the different classes, and is in fact essentially the same as the "current weight" index number discussed at the close of the previous section.

Digitized by Google

In order to avoid confusion, the author recommends that the "deflation" or "current weight" index number described in (5) be used only for special comparisons by those familiar with index numbers. The needs of most users will be cared for satisfactorily by the "constant weight" or "price-level" index number, and the volume of sales series described in (4). They must be told that the latter cannot be deduced from the former and the record of sales at current prices, because of the effects of changing weights, but no advantage is gained by giving them the exact index number which may be determined as described in (5).

The methods for determining the "price level" and the "deflation" index numbers on the basis of sales by classes are shown in the following example.

EXAMPLE

Determine the price-level index number on the basis of recent experience, the sales volume in each year at the price level of the first year, and the deflation index number, given the following data:

	Hardware		Timber Products		All Items
	Sales	Price Index	Sales	Price Index	Sales
1914.....	$60,000	100	$40,000	100	$100,000
1919.....	200,000	200	120,000	120	320,000
1924.....	300,000	150	150,000	100	450,000

Price-Level Index Number: At 1914 prices, the $300,000 worth of hardware sold in 1924 would have been worth $200,000, while the $150,000 worth of timber products would have had the same value. In accordance with the second method of solution of the typical problem discussed at the beginning of this chapter, then, we can combine the price series for hardware and timber products by weighting the former 4 and the latter 3, on the basis of relative proportions of value at base year prices. The values of the combined price-level index are therefore:

1914	100
1919	165.7
1924	128.6

Sales Volume:

$$\text{1919 sales at 1914 prices} = \frac{200{,}000}{2} + \frac{120{,}000}{1.2} = 200{,}000$$

$$\text{1924 sales at 1914 prices} = \frac{300{,}000}{1.5} + \frac{150{,}000}{1} = 350{,}000$$

Digitized by Google

The index number of volume of sales is therefore:

1914	100
1919	200
1924	350

"DEFLATION" INDEX NUMBER:

For 1919, 1919 actual sales divided by 200,000, volume at 1914 prices, = 160

For 1924, 1924 actual sales divided by 350,000, volume at 1914 prices, = 128.6

ADDITIONAL EXAMPLES

1. Determine the price-level index number, 1922 as base, on the basis of recent experience, the sales volume in each year at the price level of the year 1922, and the deflation index number (1922 as base), given the following data:

	COTTON GOODS		WOOLEN GOODS		COTTON AND WOOLEN GOODS
	Sales	Price Index	Sales	Price Index	Sales
1913.....	$2,000,000	80	$3,000,000	60	$5,000,000
1922.....	3,000,000	100	5,000,000	100	8,000,000
1925.....	5,000,000	125	4,800,000	80	9,800,000

2. A certain company sold $700,000 worth of electric lamps in 1924, and $1,400,000 worth of storage batteries. If the price index of lamps, with 1914 as base, was 80 in 1923 and 70 in 1924, and the price index of storage batteries, with 1914 as base, was 150 in 1923 and 140 in 1924, find the combined price-level index number, 1914 as base, for 1923 and 1924.

3. Discuss from various angles the changes in the cost of clothing in a certain prison, if the prices per unit and the annual quantities used per 100 prisoners were as indicated in the following table:

	PRICES PER UNIT			QUANTITIES		
	1921	1923	1925	1921	1923	1925
Suits.....	$10.00	$12.00	$11.00	50	40	45
Shoes (pairs).....	2.00	2.00	2.00	200	300	300
Shirts.....	3.00	2.50	2.00	300	350	400

Third complication: More than two series. In all the examples discussed so far, the number of items included in the index number has been small — usually two, and never over four. The arithmetic has thereby been much reduced, and the essential mathematical points brought into sharper relief. In the problems that actually arise, however, the number of items (that is to say, of series) is almost always greater than four and in some cases would be so large as to make the computations prohibitive, if all items were

Digitized by Google

used. In such cases it becomes necessary to select important or typical items as the basis for the numerical work, and the question arises, how many such items should be used?

From the point of view of the accuracy of the index number in representing conditions, it is desirable to include as many series as available records permit. A single series is likely to have peculiarities which are not common to the group. If only a small number of series are selected, several of them may happen to depart in the same way from the general trend, and hence give rise to an incorrect composite. If a relatively large number of series is included, however, the probability of the reliability of their average, as a measure of the tendency of the whole group, is increased. This point will be clearer after the discussion of Probability in later chapters.

On the other hand, from the point of view of the person who is assembling the information, it is desirable to use only a few series, if the number can be reduced without too great a loss of reliability. Each additional series represents additional labor and probably expense, both in gathering the facts and in carrying through the computation.

The practical problem, then, is to balance these two considerations in accordance with the requirements of the task in hand, and to include enough series to allow for all the really different tendencies that occur, but not to include so many as to make the work unwieldy. Serviceable price index numbers in limited fields have been constructed on the basis of ten series, but more frequently as many as 40 or 50 series, or more, have been found necessary.

Fourth complication: Changes in character of items. One of the most annoying difficulties in index number construction remains to be considered. It often happens, particularly when we try to find price index numbers of manufactured products, that many of the articles sold at the earlier date are no longer being produced at the later date, their places being taken, more or less directly, by new articles. If we should attempt, for instance, to measure the average change in prices paid for firearms between the Civil War and the World War, we should find no dealer quoting prices for weapons

Digitized by Google

of the Civil War type except as antiques or junk, while it would be still less possible to state a fair price for what a high-powered breech-loading rifle would have cost in Civil War days if the design had been known.

The example just cited is of course an extreme instance. But in the case of most complex manufactured goods, few items remain active for as long as ten years without substantial changes in design. It is therefore impossible to determine valid price index numbers, covering the entire class of items, by the methods outlined in the section "First complication — Variation in quantities." The difficulty in such cases is more serious than those due to change of weights, in that there are no prices for some items at the first date and no prices for other items at the second date. The only practical method of meeting the difficulty in many cases seems to be some form of the "chain method" discussed in the next section.

The chain method of construction. The fundamental idea of the chain method of constructing index numbers is one that would strike almost any one who began to build index numbers. If prices are found to have increased 10 per cent from 1914 to 1915, and 20 per cent from 1915 to 1916, it is fairly obvious that we might expect the price index number for 1916 with 1914 as base to be 110×1.20 or 132, or in other words, the total percentage of increase is 32. (It is to be hoped that no student at this stage of progress will try to obtain the total percentage of increase by adding 10 and 20!) The procedure suggested is undeniably satisfactory if the 10 and the 20 per cent are computed on the basis of the same items with the same quantity weights. In the practical problems discussed in the preceding section, however, the list of active items is continually changing, and may well be different at each of the three dates. In such cases the two percentages of change must be based on different lists of items, and it is necessary to consider carefully the conditions under which the final results will be reliable.

The argument will be clarified by discussion of two numerical examples, in one of which the chain method is clearly equivalent to the methods shown previously, and in the second of which such comparison is impossible.

Digitized by Google

EXAMPLE 1. Find by the chain method the index number of prices in 1925, if the prices were as indicated in table.

	QUANTITY	1914	1920	1925
Item A	200	$2.00	$5.00	$3.00
Item B	100	3.00	4.00	3.80

As in first method of first paragraph,

$$\text{Value 1914} = 2 \times 200 + 3 \times 100 = 700$$
$$\text{Value 1920} = 5 \times 200 + 4 \times 100 = 1400$$
$$\text{Value 1925} = 3 \times 200 + 3.8 \times 100 = 980$$

We have, then,

index number of second date with respect to first $= \frac{1400}{700} = 200$ per cent
index number of third date with respect to second $= \frac{980}{1400} = 70$ per cent
chain index number of 1925 with respect to 1914 $= 200 \times .70 = 140$

But obviously the index number of 1925 with 1914 as base is $\frac{980}{700} = 140$ per cent, agreeing with value just found.

The second method of the first paragraph of this chapter throws more light on what we usually want to do in the case of chain index numbers.

Per cent changes, 1914 to 1920, item A, 150; item B, 33.3
1920 to 1925, item A, -40; item B, -5
Per cent of values, 1914, item A, 57.1; item B, 42.9
1920, item A, 71.4; item B, 28.6

The average per cent changes are therefore:

$$1914 \text{ to } 1920,\ .571 \times 150 + .429 \times 33.3 = 99.9$$
$$1920 \text{ to } 1925,\ .714 \times (-40) + .286 \times (-5) = -29.99$$

The chain index number is therefore $199.9 \times (1 - .2999) = 140$, as before.

It should be noted that the value weights applied to percentage increases are always based on values at the earlier of the two dates. It would be wrong in the preceding example, for instance, to weight the percentage changes from 1920 to 1925 by .571 and .429, the 1914 percentages of value.

EXAMPLE 2. Find by the chain method the index number of prices in 1925, with 1914 as base, if the prices and quantities are as indicated.

	1914		1920		1925	
	Quantity	Price	Quantity	Price	Quantity	Price
Item A	100	$1.00	50	$1.20	0	
Item B	20	5.00	30	5.00	45	$6.00
Item C	0		10	10.00	20	8.00
Item D	0		100	0.50	1000	0.25

The index of price change from 1914 to 1920 is clearly

$$\frac{100 \times 1.20 + 20 \times 5.00}{100 \times 1.00 + 20 \times 5.00} \times 100 = 110$$

and the index of change from 1920 to 1925 is

$$\frac{30 \times 6.00 + 10 \times 8.00 + 100 \times 0.25}{30 \times 5.00 + 10 \times 10.00 + 100 \times 0.50} \times 100 = 95$$

Digitized by Google

The chain index number of 1925 prices with respect to 1914 prices is therefore $110 \times .95 = 104.5$.

A second method of determining this number would be to use the weights at the second date rather than the first. By this method, the index of change from 1914 to 1920 is 105; from 1920 to 1925, 73.5; and the combined index is 77.2. This result is widely different from the first because of the different weight given to the reduction of price of item D by the two methods. As the introductory price of a new article is often merely experimental, the first method, which automatically gives a small weight to a new item, is usually preferable to the second.

A third method would be to use as weights the averages of the quantities in the two years. As the reader may readily verify, the results are:

Index of price change from 1914 to 1920, 107.5
Index of price change from 1920 to 1925, 78.8
Resulting 1925 index, with 1914 as base, 84.7

This method has no logical argument to recommend it, except that it is a kind of compromise. Whether the purpose of the index is to measure bona fide changes in the price level or to reduce sales to a constant price level, the best method is ordinarily the first of the three, that is to say, using as weights the quantities in the earlier year. We thus give least weight to the exceptional prices in the early years of a new item. The resulting index does not reflect in full the price reductions which are almost certain to occur as production of a new item is put on a quantity basis, and is therefore more nearly comparable to price indices of similar classes of items in which no change in the list of items sold has occurred.

The following suggestions as to the use of the chain method are based on the author's experience and the algebraic analysis of the next section:

(1) As the chain method of construction is not as satisfactory as methods explained earlier, it is desirable to limit its use to a homogeneous class of items whose price changes are more or less the same. The index numbers for these classes can then be combined by the methods explained earlier.

(2) It is desirable to minimize the danger of inaccuracies due to "chaining" by using as few links as practicable. Change the list of items and weights every three or four years rather than every year, unless the number of items active throughout a period is thereby unduly reduced.

Algebraic discussion of chain index numbers. It is clear from the first example above that the chain method of construction can-

Digitized by Google

not be questioned when the quantity weights remain constant and the list of items remains unchanged. But the second example makes it clear that when the list of items and the weights are different at different dates, different variations of the chain method give different results. Let us try, by stating the problem algebraically, to throw some light on the problem of telling whether the chain method will give reliable results in a particular case.

Let us suppose, for the sake of illustration, that five articles of the class under consideration were on sale at the first date, that one of these was no longer active at the second date, but two new articles were put into production, and that at the third date two more of the original list had become inactive, and three new articles introduced. Let us denote by p_1, p_2, p_3, p_4, p_5 the prices and by q_1, q_2, q_3, q_4, q_5 the quantities at the first date. Let article corresponding to the subscript 1 be the one that becomes inactive. Then the prices at the second date will be $p_2', p_3', p_4', p_5', p_6', p_7'$ and the quantities $q_2', q_3', q_4', q_5', q_6', q_7'$. Similarly the prices at the third date will be in this notation $p_4'', p_5'', p_6'', p_7'', p_8'', p_9'', p_{10}''$ and the quantities $q_4'', q_5'', q_6'', q_7'', q_8'', q_9'', q_{10}''$.

According to the procedure recommended in the preceding section, the best chain index for the third date with respect to the first is:

$$\frac{q_2p_2' + q_3p_3' + q_4p_4' + q_5p_5'}{q_2p_2 + q_3p_3 + q_4p_4 + q_5p_5} \times \frac{q_4'p_4'' + q_5'p_5'' + q_6'p_6'' + q_7'p_7''}{q_4'p_4' + q_5'p_5' + q_6'p_6' + p_7'q_7'} \times 100.$$

An index might be derived by the methods of earlier sections, on the basis of average quantities at the first and second dates and articles 4 and 5, which remained active throughout the period. The index for the third date with respect to the first on this basis is:

$$\frac{(q_4 + q_4')\, p_4'' + (q_5 + q_5')\, p_5''}{(q_4 + q_4')\, p_4 + (q_5 + q_5')\, p_5} \times 100.$$

This latter index and the chain index will be about the same if the fourth and fifth items are much more important than the others, or if the various changes in prices and quantities happen to be such as to produce such a net result. In most cases, however, in which it is proposed to use the chain process, items such as the fourth and fifth which are active throughout the entire period are either relatively unimportant or characterized by distinctly different price changes

Digitized by Google

than the remainder of the items. In such cases most people who have to use the results feel more confidence in a product, as found by the chain method, of two percentages of change, each broadly based on a representative list of items, than on a single percentage of change based on relatively few items.

The student is urged to note carefully the formula for the chain index as given above, and to state the corresponding formula when he is considering the acceptability of a chain index for some particular purpose. It should be noted especially that the numerators and denominators in every fraction are sums of terms of the type pq. As was shown in the first section of this chapter, any correct index constructed as the average per cent of change can be reduced to the Σpq form, and therefore the formula as written indicates the only correct method. Certain discussions which the reader may discover fail to do justice to the chain method in that they construct the separate quotients by averaging percentages of change for different items without proper attention to the weights.[1]

In view of the uncertainty which necessarily characterizes the chain method, the reader will probably try to avoid it if possible. As the chain method is essential only when the list of items changes, every effort should be made to estimate prices a year or two further when items pass from the active list, and to connect new items with old which they more or less replace. In the latter case, it may be necessary to multiply prices of the replacing item by a factor to allow for increased efficiency. If such connection can be established, so that the two items can be combined in one price series, it may make chaining unnecessary, or at least reduce the number of links.

EXAMPLES

1. A conspicuous example of prices which can be summarized as an index number only by some form of the chain method is provided by stocks on the New York Stock Exchange. It is recommended that the student select 10 industrial stocks, choosing from among those which were most active in some week five years ago. Then let him note the prices for these same stocks one year, two years, etc., later, and check up the records to make sure that the stocks chosen have not been affected by mergers, stock dividends, or other

[1] For example, note on page 3 of United States Bureau of Labor Statistics Bulletin, *Index Numbers*, by S. W. Mitchell. In general, this Bulletin is the best available treatise on index numbers.

Digitized by Google

changes which really make the stocks now quoted something different than what he started with. Then let him try to make adjustment for these changes, as suggested at the end of the preceding section, and construct an index number, using the chain method if necessary.

2. Compute the index number of the price of cable in 1922 and 1925, with 1919 as base, given the following information as to quantities and prices of the various types, as indicated by their respective code-numbers.

Type	1919		1922		1925	
	Quantity	Price per ft.	Quantity	Price per ft.	Quantity	Price per ft.
C–21.....	20,000 ft.	$20.00	10,000 ft.	$15.00	0	
C–22.....	30,000 ft.	25.00	40,000 ft.	24.00	20,000 ft.	$20.00
C–23.....	10,000 ft.	30.00	25,000 ft.	32.00	50,000 ft.	30.00
C–24.....	0		1,000 ft.	40.00	10,000 ft.	32.00

Use of the geometric mean. Before the critical discussion of index numbers had progressed as far as it now has, some people attempted to find index numbers for a group of items by reducing the series of prices for each item to a series of relatives, with the same year as base for all items, and then finding the arithmetic average of these relatives for a given year as the index number for that year. As we have explained, this is sound only when each series is weighted by its proportion of values in the base year. If the series are combined with equal weights, it will be found that widely different results are found according to what year is taken as base. In order to avoid the difficulties to which this use of the equal-weighted arithmetic mean led, some writers have used or advocated the geometric mean. The two chief advantages of this mean are:

(1) Essentially the same series of index numbers is obtained by taking the relative number of the geometric means of the quantities themselves as by taking the geometric means of the relative numbers.

Algebraically, if the series are

$$a_1, a_2, a_3, \ldots a_n$$
$$b_1, b_2, b_3, \ldots b_n$$
$$c_1, c_2, c_3, \ldots c_n$$
$$\ldots\ldots\ldots\ldots \text{(to } r \text{ series)}$$

Digitized by Google

then the relative number of the geometric means, first year as base, is

$$\frac{\sqrt[r]{a_n \times b_n \times c_n \ldots}}{\sqrt[r]{a_1 \times b_1 \times c_1 \ldots}} \times 100$$

and the geometric mean of the relative numbers, first year as base, is

$$\sqrt[r]{\frac{a_n}{a_1}}\ \sqrt[r]{\frac{b_n}{b_1}}\ \sqrt[r]{\frac{c_n}{c_1}} \ldots \times 100$$

which is obviously the same as the preceding expression.

(2) The same ratio of index numbers for two dates is obtained no matter what year is taken as base, if the geometric mean is used in averaging the relative numbers for each date. For the index number of the first year, nth year as base, is

$$\frac{\sqrt[r]{a_1 \times b_1 \ldots}}{\sqrt[r]{a_n \times b_n \ldots}} \times 100,$$

and the index for the second year, nth year as base, is

$$\frac{\sqrt[r]{a_2 \times b_2 \ldots}}{\sqrt[r]{a_n \times b_n \ldots}} \times 100,$$

and the ratio of these two indexes is clearly independent of n, that is to say, is the same whatever year is chosen as base.

In view of these two properties, the geometric mean is to be preferred to the arithmetic when the index must be constructed as an average of relatives with no natural system of weights available. If, for instance, we want to average several measures of progress for an individual, and the units of the different series are pounds, inches, and academic grades, there is no common unit by means of which these quantities can be combined into a total. Use of the geometric mean avoids the bias which will result from the choice of a base year for the relatives, if an arithmetic average is found. If a base year is chosen in which one of the quantities was abnormally low, it will be found that the arithmetic mean method leads to the same result as giving that quantity greater weight than the others.

Digitized by Google

The danger of such an unintentional bias is avoided if the geometric mean is used.

Applications of index numbers. Index numbers are the most useful tool which the economist has. Among the problems in which they are useful, attention is called to the following, some of which have been discussed previously:

1. Determination of the level of wholesale prices, of the cost of living, of money wages, of real wages (i.e., money wages divided by cost of living), or other similar variable.
2. Adjustment of wage scales to allow for changes in living costs. For this purpose, it is essential to secure an index of living costs which really applies to the group whose wages are being revised; for example, Bradstreet's index number of wholesale prices does not show what changes should be made in the salaries of Harvard professors.
3. Elimination of the effects of price variation from historical series of sales, purchases, pay rolls, or similar quantities.
4. Measuring volume of production or output per employee by summarization of data expressed in physical units.
5. Measuring changes in efficiency by comparing an index of shop-costs with a composite index of materials used, wage rates, salaries, and miscellaneous overhead.

Other problems of a more theoretical nature could be cited. Those listed are enough to indicate that, in the struggle which is now going on to transform economics from a semi-literary and philosophical speculation to a science based on observed and analyzed facts, the index number is one of the "Big Berthas" — or rather, is the standard type of artillery for laying down the barrage beneath whose protection the trenches of the opposition are captured and consolidated.

Summary of Chapter VI. Business men and economists frequently desire to express in a single figure the combined effect of price changes in two or more articles.

One method of accomplishing this is to compute the ratio of cost

Digitized by Google

of quantities used at the prices in effect at the second date to cost of same quantities at the prices of the earlier date. Another method, which gives the same result, is to compute the weighted average ratio of price increase, the weight for each article being the cost of the quantity used at the prices of the earlier year.

The ratio resulting from either of these methods is a simple illustration of a price *index number*. An index number is an average or composite of two or more historical or geographic series, which indicates the variations from time to time, or from place to place, in the price, cost, or wage level, or in some other economic variable.

If the quantities used are not the same at the second date as at the first date, several different methods for computing the index number may be used. The choice of method depends in part on the purpose for which the number is to be used in the particular problem. It is often satisfactory to compromise by using the average of the weights or taking the geometric mean of the index numbers resulting from the two sets of weights.

If more than two dates are involved, the question of which weights to use becomes more complicated. If the purpose is to show fluctuations in the price level, it is usually best to base the weights on the average experience of recent years. If the purpose is to remove the effect of price variation from a record of sales (or similar variable), it is best to break the total up into groups such that the articles in any group have varied somewhat similarly in price, and then divide the sales in each group in each year by the index number for that group based on quantities sold in that year.

The number of series used in constructing an index number should be large enough to cover the different tendencies which may exist, but not so large as to cause undue difficulty in the collection and preparation of the data.

If the list of items sold or manufactured has changed during the period under consideration, it is often necessary to use the *chain method* of construction. This method consists of multiplying together an average ratio of change from the first date to the second and an average ratio of change from the second date to a third date, in order to obtain the ratio of change from the first date to the third. The best weights to use in the first of these ratios are the quantities

Digitized by Google

of the first date, and in the second ratio the quantities of the second date. The chain method should be limited to classes of items whose price changes are more or less the same, and the number of links should be kept as small as practicable without unduly curtailing the number of items active throughout the period covered by one link.

The *geometric mean* of the relative numbers is preferable to the equal weighted arithmetic mean, in case things of a radically different nature are being summarized. This preference arises from the fact that the same index is derived from the geometric mean of the quantities themselves as from the geometric mean of the relative numbers, and that the same ratio of index numbers for two dates is obtained no matter what year is taken as base.

Index numbers are an indispensable tool in a wide variety of problems treated by the economist or business statistician.

Digitized by Google

CHAPTER VII

MEASURES OF DISPERSION

In Chapter IV we showed the importance of the frequency distribution as a presentation in full of facts in common life often summarized by a rough average, or as a complete description of the magnitude of a certain characteristic in a group of individuals, as a basis for scientific discussion. In Chapter V, we discussed the five principal averages, each considered as a device for presenting in a single figure the position of a frequency distribution, or in other words, the value of a characteristic which can fairly be regarded as typical from some standpoint. In this chapter we shall take up an intermediate position, and discuss measures which are used to give a supplementary summary of a frequency distribution.

Let us analyze how fully an average, or even two or three averages — say the arithmetic mean and the median — taken together summarize a frequency distribution. Consider, for example, the frequency distribution of the length of 326 ears of corn for which various averages were found in Chapter V. Though the mean, mode, and median are clearly grouped at 9, yet 2 ears are 12 inches long and 1 only 3 inches long, and the number either 8 or 10 inches long is more than half as large as the number at 9. In order to give an idea of the "spread" of the items without the fulness of detail of the frequency distribution, it has been found desirable to use what are known as "measures of dispersion." They have been used frequently in connection with distribution of incomes, students' grades, prices, errors, and anthropometric measurements.

The range. One of the commonest of these measures is the range, the interval from the smallest to the largest item in the distribution. This measure is in common use, though not favored scientifically. We say, for instance, that the age of freshmen varies from 15 to 29; that the income of doctors varies from nothing to $50,000 a year, etc. It is at once obvious, however, that the range may be very much affected by the presence or absence of a few extreme cases;

Digitized by Google

another year, for instance, the age of freshmen might vary from 16 to 23, with no real difference in the spread for the two years except that a boy of 15 and a man of 29 were included the first year but not the second. In common speech, we avoid this difficulty by barring exceptional cases; but such a term is differently interpreted by different people. It is therefore desirable to use the "range" only in a rough way, and develop other measures for accurate purposes.

The quartile deviation. One of the simplest measures of dispersion is based on the determination of the *quartile points*, that is to say, the points which divide the distribution into quarters, as the median divides it into halves. The value of the quantity at the quartile point which one comes to first, starting with the lower values, is called the *lower quartile*, and is usually denoted by the symbol Q_1; the median is really the second quartile, but is not usually so called; the value of the quantity at the last quartile point is called the *upper quartile*, and is denoted by Q_3. The four sections of the distribution are also referred to as the first, second, third, and fourth quartiles, starting as before with the numerically smallest values.

The difference between the values of Q_1 and Q_3 is called the *interquartile range;* one half of this the *semi-interquartile range* or the *quartile deviation.* For reasons which will be developed in the chapter on probability, it is also sometimes called the probable error; it is recommended, however, that this phrase be restricted to cases which involve some idea of probability or error.

The computation of the quartile deviation is illustrated by the following example, the same as was used for several other purposes in earlier chapters. The method of computation is essentially the same as that used in computing the median in Chapter V.

Length of ear.....	3	4	5	6	7	8	9	10	11	12
Number of ears.....	1	1	3	14	25	62	119	80	19	2
Cumulative No.....	1	2	5	19	44	106	225	305	324	326

One fourth of 326 is 81.5; Q_1 is therefore

$$7.5 + \frac{81.5 - 44}{62} = 8.10$$

Digitized by Google

and similarly

$$Q_3 = 9.5 + \frac{244.5 - 225}{80} = 9.75$$

As found before, the median is 8.98. The interquartile range is 1.65, and the quartile deviation is then .82. It is to be noted that the median is distinctly nearer to Q_3 than to Q_1. This difference shows the departure from symmetry of this distribution.

In the case of a series in which the items are not grouped into classes, as they were in this example, it is convenient to use formulas analogous to the $\frac{(N+1)}{2}$ formula used for the median. These are, that the first quartile is the value of the quantity for the $\frac{(N+1)}{4}$ item in the rank list, and the third quartile is the value of the quantity for the $\frac{3(N+1)}{4}$ item.

The computation of the quartile deviation for ungrouped items is shown in the following example. The salaries of 10 men are \$2000, 2500, 2800, 3000, 3200, 3300, 3500, 3800, 4500, 5000. Find the quartile points and the quartile deviation. Here $N = 10$, the first quartile is the salary of the $\frac{11}{4}$ man, and the upper quartile that of the $\frac{33}{4}$ man, i.e., Q_1 is the salary three fourths of the way from that of the second man to that of the third man, and Q_3 is the salary one fourth of the way from that of the eighth man to that of the ninth man. We have therefore:

$$Q_1 = 2725, \quad Q_3 = 3975, \quad Q = \frac{Q_3 - Q_1}{2} = 625.$$

Mean deviation, or average deviation. For symmetrical distributions, the quartile deviation is really the median of the deviations about the mean, taken without regard to sign. A second measure of dispersion is the Mean or Average Deviation, that is to say, the *arithmetic mean* of the deviations, taken without regard to

Digitized by Google

sign about some value, usually the mean or the median. As this measure can be computed without arranging the items in a rank list or frequency distribution, it is in some ways easier to compute than the quartile deviations, and in any case is not difficult.

The computation of the mean deviation about the mean for the example (10 salaries) used above, is as follows:

	Deviation from assumed mean 3200		Deviation from true mean 3360	
2000	−1200		−1360	
2500	− 700		− 860	
2800	− 400		− 560	
3000	− 200		− 360	
3200		100	− 160	
3300		300	− 60	
3500		600		140
3800		1300		440
4500		1800		1140
5000				1640
	−2500	4100	−3360	3360
		−2500		
		1600		

$$\text{True mean } 3200 + \frac{1600}{10} = 3360$$

$$\text{Total deviation about true mean} = 3360 + 3360 = 6720$$

$$\text{Mean deviation} = \frac{6720}{10} = 672$$

The average deviation may also be computed by using the assumed mean 3200, as follows. We first note that the total deviation about 3200 is 2500 + 4100 = 6600. When we change the reference point from 3200 to 3360, we increase the deviations of five items (2000–3200 incl.) by 160, decrease the deviations of 4 items (3500–5000 incl.) by 160, and change the deviation of one item (3300) from 100 to 60. The net change is therefore an increase in the total deviation of 160−40 = 120; the total deviation about 3360 is therefore 6720, as found by the method shown above. This method is worth considering as an introduction to the computation of the average deviation for frequency distributions, shown below.

An important property of the mean deviation is that *it is a minimum about the median;* this will be proved in a later section.

Digitized by Google

In spite of this property, the mean deviation is more often taken about the mean than about the median. The reason for this is probably that the total deviation about the mean is already nearly computed in the process of determining the mean by the deviation method. The only additional step required is applying the correction to transfer the center of reference from the assumed to the true mean.

As this correction is a little troublesome, let us consider in detail its application to the frequency distribution of the length of 326 ears of corn. In finding the average length of ear in Chapter V, the following table was used:

Length	Frequency — No. of Ears	Deviation from Assumed Mean 9	Deviation Times Frequency	
3 in.	1	−6		− 6
4	1	−5		− 5
5	3	−4		−12
6	14	−3		−42
7	25	−2		−50
8	62	−1		−62
9	119	0		0
10	80	1	80	
11	19	2	38	
12	2	3	6	
Total	326		124	−177

We note that the product "deviation times frequency" was found to be 124 positive and 177 negative, with "9 inches" taken as assumed mean. Disregarding sign, we then have total deviation, uncorrected, equal to 301. In this computation, however, the assumption was made that all the ears in a given interval could be treated as if they were concentrated at the mid-point of the interval. If we are taking deviations about a point outside of the interval, this assumption is in accord with the uniform distribution of ears within an interval, as the total deviation for two ears at equal distances on either side of the mid-point is the same as if they were at that mid-point. If we are taking total deviation without regard to

Digitized by Google

sign about the mid-point itself, however, the assumption of concentration leads to a zero total deviation, which is obviously incorrect. Consider all the ears in the upper half of the interval; on the average they are distant from the mid-point by one fourth of the interval, if they are numerous enough to justify using a uniform distribution as an approximation to the actual. Similarly the average deviation for ears in the lower half of the interval is one fourth of an interval. The total deviation for all ears in the interval, ignoring, as we must for present purposes, differences in sign, is therefore the product, number of ears in interval times one fourth of an interval.

In the numerical case we are considering, therefore, the uncorrected total deviation should be increased by $119 \times \frac{1}{4} = 30$, and becomes 331.

The next step is to transfer from the assumed mean "9" to the true mean, found in Chapter V to be 8.84. When we move the reference point from 9 to 8.84, the deviations for all ears longer than 9 are increased .16 of an inch; the deviations for all ears shorter than 8.84 are decreased by .16 of an inch; and on the average the deviations for ears between 8.84 and 9 inches in length are unchanged, if the ears are frequent enough to justify the use of a uniform distribution as an approximation to the actual. If we had a uniform distribution, the change of reference point would amount merely to finding deviations for all points in a line-segment, about one end of the segment rather than the other. It is obvious by symmetry that two such totals must be the same.

The necessary correction, then, is to *add*

$$.16 \text{ times } \left(2 + 19 + 80 + \frac{119}{2}\right)$$

and *subtract* .16 times $(106 + 119 \times .34)$, a net *addition* of 2. The corrected total deviation is therefore 333, and, dividing by 326, we find that the average deviation is 1.02.

The usual method of computing this correction is to assume all the items in the interval concentrated at its mid-point. We then have $101 + 119$ $(= 220)$ items whose distance from the true mean is greater than their distance from "9," and 106 whose distance is

Digitized by Google

less. We then have a net number 114 whose distance is greater, in each case by .14. The correction to be added is therefore $114 \times .14$ or 16. This gives 317 for the sum and .97 for the mean deviation, a result distinctly too small. The important thing given incorrectly in this case is the total deviation for items in the interval in which the mean lies; if the interval is a considerable proportion of the entire range, this error is important. If the number of intervals is large, and each interval narrow, items in the interval in which the mean lies may be ignored in computing the correction.

An absurd result of this usual method of computing the correction is the fact that, contrary to the minimum property previously stated, the mean deviation about the median will always be greater than the mean deviation about the mid-point of the class interval in which the median lies.

If the items are actually concentrated at the mid-point of the interval, this usual method is correct. In the general case, however, the items are distributed uniformly over each interval, or that is the most reasonable assumption to make.

General formula for mean deviation of a frequency distribution. It is advisable to state in general terms exactly the steps to be taken in computing the mean deviation in this case:

First, compute or note the sum, say S, of the deviations without regard to sign, about some assumed mean A on the assumption that all items are concentrated at the mid-point of the interval in which they lie. This sum is a by-product of the computation of the mean by the deviation method.

If the true Mean M (or the true Median M'), about which the mean deviation is being computed, does not lie in the same interval as A, it is necessary to find S', the sum of the deviations without regard to sign about A', the mid-point of the interval in which M (M') does lie. S' is easily seen to be S minus the product of the distance from A to A' into the net number of items for which distance from A is greater than distance from A'.

Let w be the width of the interval in which A' and M lie, n_1 the number of items below this interval, n_2 the number in that interval, and n_3 the number above that interval, and c the amount by which

Digitized by Google

M exceeds A (negative if A exceeds M). Then as in the example worked out, S (or S') must be increased by $\frac{w}{4} n_2$ to allow for uniform distribution over the central interval rather than concentration at the center. The total of deviations so far is S (or S') $+ \frac{w}{4} n_2$.

If now we move the reference point from A to M, we increase the deviations of $n_1 + \frac{n_2}{2}$ items each by c units, and decrease the deviations of $n_3 + (\frac{1}{2} - \frac{c}{w}) n_2$ items by the same amount. The net increase is then the difference of these, or

$$\left\{ (n_1 + \frac{n_2}{2} - [n_3 + (\tfrac{1}{2} - \frac{c}{w}) n_2] \right\} c$$

which may be reduced to $(n_1 - n_3 + \frac{c}{w} n_2)c$, and the same expression will be found to result if M is below A and c is taken negative. The total deviation is therefore:

$$S + \frac{w}{4} n_2 + (n_1 - n_3 + \frac{c}{w} n_2)\, c, \qquad \text{(Formula VII–1–a)}$$

and the mean deviation is this total divided by $n_1 + n_2 + n_3 = N$. Another way of stating the last formula is to write it:

$$S + (n_1 + n_2 - n_3)\, k + \frac{n_2 \left(k - \frac{w}{2}\right)^2}{w} \text{ if } M \text{ is } k \text{ above } A,$$

(Formula VII–1–b)

or $$S + (n_3 + n_2 - n_1)\, k + \frac{n_2 \left(k - \frac{w}{2}\right)^2}{w} \text{ if } M \text{ is } k \text{ below } A.$$

(Formula VII–1–c)

Note that k is always taken positive.

If k is nearly equal to $\frac{w}{2}$, this last formula is nearly that ordinarily used, as the last term is negligible. If on the other hand k (i.e., the

Digitized by Google

numerical value of c taken without regard to sign) is small, the last term of the first formula is negligible. The importance of the $\frac{w}{4}n_2$ term, usually neglected, depends on the relative magnitude of n_2 compared with N, and of w compared with entire range.

The mean deviation is a minimum about the median. This may easily be proved. Consider first the case in which the number of items is odd. In this case the median is located exactly at the $(n+1)/2$ item, and n times the mean deviation about the median is the sum of the lengths from each item to this middle item, if we take each item as given a position on a line according to the value of the variable for that item. Consider how this sum is changed as we shift to the right or left the point to which the lengths are measured. There are $(n-1)/2$ items on each side; if we move to the right, say, the lengths to items on the right are diminished, and the lengths to items on the left are increased, each by the same amount. But as the median item is also on the left of the new point to which lengths are measured, there is one more item to the left than to the right, and the increase is therefore greater than the decrease. If we continue to move to the right until we cross over the position of another item, there will then be 3 more points to the left than the right, and the increase for each unit of movement will then be greater. It is therefore clear that the sum of the lengths to the original position (i.e., to the median) is less than to any point on the right, or, since the same argument would hold for moving to the left, the sum is less than it would be to any other point.

If n is even, there are $n/2$ items to the right and $n/2$ items to the left, and moving to the right or left does not change the sum until we cross over the position of one item. When we do, on practically the same argument as above, the sum is increased. We can see then that the sum is a minimum for the median in this case also, in the sense that no other corresponding sum is less, and that the sum for any point between or at one of the two middle items is less than for any other point. Therefore, with this understanding, we can say that the mean deviation (i.e., the sum divided by n) is a minimum about the median.

Digitized by Google

The standard deviation. There is a third valuable measure of dispersion, which on first sight appears more abstruse than those just considered, and in fact is less obviously valuable. In the event, however, it will be seen to yield itself much more readily to certain theoretical discussions, and to be desirable for use in a large number of fields. It is known as the standard deviation, and is usually denoted by the Greek letter σ (sigma).

It was obvious at the beginning of the discussion that we could not obtain a measure of dispersion merely by algebraic addition of the deviations, as the result would be zero. We obtained the mean deviation by ignoring signs and adding, and then dividing by number of items. We might also get rid of the signs by squaring, and that is what is done in finding the standard deviation. The rule is: Square each deviation, multiply by its frequency, add the products; divide by the total number of items, i.e., by the sum of the frequencies; and find the square root of the quotient. The formula is

$$\sigma = \sqrt{\frac{\Sigma f(x - \bar{x})^2}{N}} \qquad \text{(Formula VII–2)}$$

where the symbol Σ indicates "sum for all items."

The method of computation, in the case of an ungrouped series, is shown below for the same example as was used to illustrate quartile and average deviation.

Salary	Deviation from true mean 3360	Deviation²
2000	−1360	1,849,600
2500	−860	739,600
2800	−560	313,600
3000	−360	129,600
3200	−160	25,600
3300	−60	3,600
3500	140	19,600
3800	440	193,600
4500	1140	1,299,600
5000	1640	2,689,600
	Sum	7,264,000

$$\frac{\text{Sum}}{10} = 726{,}400$$

square root of above = 852

Digitized by Google

In carrying out the arithmetic, especially as applied to a frequency distribution, it is often essential to make use of a formula correcting the value for the assumed mean to that for the true mean. This formula is, that if σ is the standard deviation about the true mean, and σ' that about the assumed mean, and c the distance between these means, then $\sigma^2 = \sigma'^2 - c^2$. The proof is as follows:

Let the distance of an item from the assumed mean be denoted by x, and its distance from the true mean by x'. Then if the true mean is c units to the right of the assumed mean, $x = x' + c$.

Squaring, we have

$$x^2 = x'^2 + 2cx' + c^2.$$

For each item, that is to say, for each pair of values of x and x' we have an equation of this sort. Adding up all these equations for the entire N items, we have: $\Sigma x^2 = \Sigma x'^2 + 2c\Sigma x' + Nc^2$. But as x' corresponds to $x - \bar{x}$ in the formula above, $\Sigma x'^2 = N\sigma^2$, and similarly $\Sigma x^2 = N\sigma'^2$. $\Sigma x'$ is the sum of the deviations about the true mean, and therefore = 0. We have therefore

$$N\sigma'^2 = N\sigma^2 + Nc^2$$

or dividing by n and transposing,

$$\sigma^2 = \sigma'^2 - c^2 \qquad \text{(Formula VII–3)}$$

It follows at once from this formula that the root mean square deviation, i.e., σ', computed about any other point than the mean, is larger than σ, or that σ about the mean is a minimum.[1]

By the use of this formula, the standard deviation for the discrete series discussed above may be computed without use of minus signs and without any subtractions. For if we take zero as the assumed mean, the correction c is the mean of the series, $\bar{x}$, and the formula for standard deviation reads:

$$\sigma^2 = \sigma'^2 - \bar{x}^2 = \frac{N\sigma'^2 - N\bar{x}^2}{N} = \frac{\Sigma x^2 - \bar{x} \cdot \Sigma x}{N}$$

[1] This formula is familiar to students of mechanics in connection with radius of gyration, and for their benefit it is worth while to state here that the standard deviation is merely the radius of gyration of the area under the frequency curve about the ordinate through the center of gravity.

Digitized by Google

The computation is:

x	x^2
2000	4,000,000
2500	6,250,000
2800	7,840,000
3000	9,000,000
3200	10,240,000
3300	10,890,000
3500	12,250,000
3800	14,440,000
4500	24,250,000
5000	25,000,000
33600	120,160,000

$$\sigma^2 = \frac{120{,}160{,}000 - 33600 \times 3360}{10}$$

$$= \frac{120{,}160{,}000 - 112{,}896{,}000}{10}$$

$$= 726{,}400 \text{ as above,}$$

and $\sigma = 852$

The method of computation for a frequency distribution will be illustrated by our old friend the ears of corn example. It is to be noted that the work consists primarily of adding another column to the form already used in finding the arithmetic mean by the deviation method. Each entry in this column is the product of the corresponding entries in the two preceding columns.

COMPUTATION OF THE STANDARD DEVIATION OF THE LENGTH OF 326 EARS OF CORN

LENGTH	FREQUENCY — NO. OF EARS	DEVIATION FROM ASSUMED MEAN 9	DEVIATION TIMES FREQUENCY	DEVIATION SQUARED TIMES FREQUENCY
3 in.	1	−6	− 6	36
4	1	−5	− 5	25
5	3	−4	−12	48
6	14	−3	−42	126
7	25	−2	−50	100
8	62	−1	−62	62
9	119	0	0	0
10	80	1	80	80
11	19	2	38	76
12	2	3	6	18
Total	326		−177 124	571

As found in Chapter V (p. 84), $c = .16$, and the true mean is

Digitized by Google

8.84. We have here $\sigma'^2 = 571/326 = 1.752$. $c^2 = .16^2 = .0256$. Therefore $\sigma^2 = 1.752 - .026 = 1.726$, and $\sigma = \sqrt{1.726} = 1.31$.

Relations between, Q, $A.D.$, and σ. The reader may note that for the above example, $Q/\sigma = .63$, and $A.D./\sigma = .78$. It is worth while to enquire whether these ratios have in general about such values, as if so they may be used as rough checks on the accuracy of the computations, and it will be possible to know roughly what the other measures are after we have computed one of the three. The following table shows the values of these two ratios, and also of $Q/A.D.$, for certain theoretical frequency distributions, and also for a number of actual distributions. The first values given are those for the so-called "normal curve," which has a basis in the theory of probability, is much used, especially in the statistics of educational tests, and will be taken up in later chapters of this book.

TYPICAL VALUES OF RATIOS $A.D./\sigma$, Q/σ, AND $Q./A.D.$

Curve or example	$A.D./\sigma$.	$Q./\sigma$	$Q./A.D.$
Theoretical:			
Normal curve	.798	.674	.845
Gamma curve, $n = 16$	.794	.667	.840
" " $n = 9$	.791	.661	.836
" " $n = 4$	.781	.642	.822
Actual distributions:			
Length of ears of corn	.78	.63	.80
Size of sections, Brown University	.68	.50	.73
Jones Statistics, I	.764	.606	.794
II	.84	.77	.91
III	.812	.688	.85
Yule, *Statistics*, chap. VIII:			
Example I	.73	.61	.84
Example II	.81	.71	.88
Example III	.80	.68	.85
Height of 1304 Scottish men	.808	.651	.806

It will be noted that the actual values of each ratio do not as a rule differ widely from the theoretical value for the normal curve. The existence of a wide discrepancy is an indication that the distribution is in some way peculiar. The second actual distribution, for instance, is J-shaped; Ex. I from Yule is derived from a distribu-

Digitized by Google

tion consisting of 38 items only. In general, we expect these ratios to be steadier as the number of items in the frequency distribution is increased.

Use of σ as a standard. The ratios of the actual deviations to the standard deviation (or to Q or to $A.D.$) are frequently used in the further analysis of a problem in place of the actual deviations themselves. If this is done, the deviations of two or more quantities, each from their respective means, are directly comparable. Consider two economic series, for instance the price of stocks, shown by an index number, and the production of pig iron. As the series stand, there is no way of comparing changes in these quantities except by a rough estimate from their graphs. If each of the deviations is divided, however, by the standard deviation based on the entire series of which it is a part, we know that a value "2" for one series represents about as unusual a condition as a value "2" for the second series. Or as another example, take two series showing the fluctuations in prices of stocks and bonds. The second series is as it stands much more stable than the former, but if we reduce each to "standardized" deviations, in the way suggested, we can begin to compare the relative changes in the two series.[1]

The quartile deviation is not so good as the standard deviation for such a purpose as this, because for unsymmetrical distributions it must be used with deviations from the mid-point between the quartiles, in order to have the results mean what they seem to. That is to say, the point for which the deviation is "1" on this scale should by the definition of quartile, have exactly one fourth of the items above it. If the deviations from the *mean* are used, and divided by Q, this will not be true for unsymmetrical distributions. The difficulty mathematically of dealing with the quartile deviation in any general way is also an obstacle to using it as a standard.

The quartile deviation is used very widely in educational statistics, with the name "probable error," for problems connected with scales for educational tests. The assumption is made in this use that the distribution is normal, and in that case the difficulty ex-

[1] Compare *Agricultural Prices* by Wallace, pp. 90, 91, 92.

Digitized by Google

plained above no longer holds. The extension of such educational methods to cases in which the distribution is not normal is made more difficult by this use of "probable error" rather than average or standard deviation.

The mean deviation might well be used for the purposes mentioned, but as a matter of fact has not been so used to any great extent.

Proportion of cases within various distances of mean. It will assist in forming an idea of the real significance of σ if one bears in mind roughly the proportion of the items in the frequency distribution which are σ units, 2σ units, etc., distant from the mean in typical distributions. The following table shows this. It can be converted roughly into Q, or $A.D.$ units by the use of the table given earlier in this chapter; for instance, the spread from 3σ to -3σ corresponds to one from 4.5 Q to -4.5 Q or 3.8 $A.D.$ to -3.8 $A.D.$ This relation is pretty exact in the case of the normal curve, and roughly correct in the case of many other curves.

PER CENT OF ITEMS MORE THAN DISTANCE INDICATED ABOVE OR BELOW MEAN

	NORMAL CURVE	"GAMMA," $n = 16$	"GAMMA," $n = 4$	EARS OF CORN
More than 3σ below mean	.135	.0006	0	.6
2σ	2.275	.82	0	4.6
1σ	15.87	15.56	14.29	14.1
Below mean	50	53.33	56.65	44.8
More than σ above mean	15.87	15.65	15.12	15.0
2σ	2.275	3.44	4.24	1.0
3σ	.135	.55	1.03	.0

It is to be noted that if the ears of corn polygon is reversed in direction, it corresponds to the general type of the Gamma curves.

For convenience in quotation, it is worth while to put the facts in the table in various other forms. For the normal distribution, for instance, 99.7 per cent of all cases fall within the interval -3σ

Digitized by Google

to 3σ, and just over 95 per cent within the interval -2σ to 2σ. Similarly 68.3 per cent of all cases fall within the interval $-\sigma$ to σ. These figures vary from curve to curve, but the approximate values are important as giving a meaning to the standard deviation. For the normal curve, using the quartile deviation, we know of course that one half of all cases fall between $-Q$ and Q; it is also true that 82.3 per cent of all cases fall between $-2Q$ and $2Q$; 95.7 per cent between $-3Q$ and $3Q$; and 99.3 between $-4Q$ and $4Q$. It is to be noted that the percentages given are exactly true only for the normal curve, a particular theoretical curve which would arise from certain simple probability problems; but other curves do not usually differ widely from this; and as a first approximation, these percentages are worth bearing in mind.

Other measures of dispersion. Other measures of dispersion which have been used are either modifications of those already discussed or variant applications of the same general ideas. The most important modification is the *coefficient of variation*, based on the standard deviation. It is obvious that we cannot understand the significance of a standard deviation of a units until we know in some way the magnitude of the quantities with which we are dealing. As the mean of the quantities shows their size in a general way, Professor Karl Pearson suggested using the per cent which the standard deviation is of the mean (in symbols, $100\sigma/\text{mean}$) as a measure of relative dispersion. This per cent is known as the *coefficient of variation*, and is frequently used, although it is not appropriate in all cases. For instance, in the case of school grades, the meaning of grades assigned by different teachers and in different schools depends on the passing grade, which is in a sense arbitrary. The mean grade will usually be low when the range from the passing grade to the highest grade is large, and therefore the coefficient of variation will be relatively large, whether the real variability is in fact large or not. The point is that the coefficient of variation is not valid in any case in which the reference point of the measurements is in some sense arbitrary. Measurements of temperature are the clearest example.

It was noted in the Preface that errors of measurement in physics

Digitized by Google

and chemistry had some of the characteristics of statistical variables, but as a matter of fact we have not drawn any illustrations from that field. Using the terminology of the preceding paragraph, we can now state the reason for this. The point is that the coefficient of variation of such a group of measurements is so small, in general, that different methods are appropriate than apply in most statistical problems.

Other measures of dispersion are similar to the quartile deviation. There are in occasional use a number of divisions of the frequency distribution similar to that into fourths expressed by the quartiles. The tertiles, for instance, the thirds of the distribution, correspond to the popular division into (a) large, average, small, or (b) short, of medium height, tall. The quintiles, the fifths of the distribution, have the advantage over the quartiles that the items about in the middle are not near a dividing line, and also that the dividing points are also deciles (tenths). A more flexible classification is into percentiles dividing the distribution into 100 equal parts. The division points on this basis include those for the quartile, quintile, and decile divisions, and suggest various measures of dispersion, such as the range from the fifth to the ninety-fifth percentile, including 90 per cent of the cases. This scheme has the advantages of the range, but avoids its instability. If it is essential to give a summary picture of a frequency distribution, and for some reason a graph cannot be drawn, probably giving the deciles is the best way, combining a considerable amount of detail with moderate compactness.

Instances of use of measures of dispersion. By using the idea of a measure of dispersion, we can sometimes state results which cannot be covered by the notion of average. For instance in comparing the ability of men and women in mental traits there are some traits for which the difference in the means is small but the standard deviation for the men is larger than for the women. The result is that much larger proportion of the men will be above a certain high standard. If this high standard means qualification for success in a profession or genius, the men will be more numerous among the geniuses — and also at the other end among the idiots.

Digitized by Google

Summary of Chapter VII. Averages, as discussed in Chapter V, summarize frequency distributions by selecting a single value of the quantity as most typical. In order to describe distributions more completely, certain *measures of dispersion* are used to show how closely the items are concentrated about the average.

The interval from the smallest to the largest item, known as the *range*, is too much affected by the chance position of the end items to be valuable except for rough purposes.

A more valuable measure is the *quartile deviation*, one half the distance between the lower and the upper quartile points. The lower quartile point is the value of the quantity such that one quarter of the items have a value less than or equal to that value; the upper quartile point is the value of the quantity such that one quarter of the items have a value greater than or equal to that value.

The *mean deviation* about the mean (or the median) is the arithmetic mean of the deviations about the mean (or median) when all deviations are taken to be positive. The mean deviation about the median is less than or equal to the mean deviation about any other point, the equality being possible only when the number of items is even and applying only to other points between the two middle items.

The formulas stated on page 142 enable one to compute the mean deviation accurately for any frequency distribution, but do not differ appreciably from the formulas usually employed if the mean is near the boundary of two intervals or if the number of intervals is large.

The *standard deviation*, the "root-mean-square" of the deviations, is computed by the formula $\sigma = \sqrt{\frac{\Sigma f (x - \bar{x})^2}{N}}$. It is more suitable for algebraic discussion than the measures of dispersion mentioned above.

As a general thing, the mean deviation is about four fifths, and the quartile deviation about two thirds, of the standard deviation. These ratios are often not valid for actual distributions based on a small number of items, but are more likely to be approximated if the number of items is large.

Digitized by Google

By relating to their respective standard deviations the actual deviations from their means of the items in two or more statistical series, we can sometimes make a comparison which would not be practicable if some such adjustment were not made.

In any frequency distribution it is unusual for more than one half of one per cent of the items to differ from the mean by as much as three times the standard deviation. Somewhat more than two thirds of the items are concentrated within the interval from one standard deviation less than the mean to one standard deviation greater than the mean.

The coefficient of variation, the per cent which the standard deviation is of the mean, may sometimes be used with advantage to show the relative variability of a particular distribution.

Tertiles, deciles, and percentiles are used to divide frequency distributions in the same way as quartiles, and can be used as the basis of other measures of dispersion similar to the quartile deviation.

The idea of dispersion, with a definite basis of measurement, makes it possible to think definitely in regard to certain problems, such as the distribution of ability, which would otherwise remain vague matters of opinion.

PROBLEMS ON CHAPTER VII

1. The salaries of professors in publicly supported institutions in the academic year 1919–1920 were as indicated by the following frequency distribution showing number in each salary group. Find the quartile, mean, and standard deviations, and the 90 per cent range. Answers: Quartile 604; *A.D.* about mean 727; about median 718; $\sigma = 913$. 90 per cent range = 3043.

Salary	Number
$250– 750	2
750–1250	4
1250–1750	50
1750–2250	302
2250–2750	628
2750–3250	552
3250–3750	372
3750–4250	271
4250–4750	127
4750–5250	94
5250–5750	45
5750–6250	10
6250–6750	1
7750–8250	1
9750–10250	1

Digitized by Google

2. The monthly wages of 1000 typical railroad workers were distributed as follows (1918 figures):

Wage.....	$10–30	$30–50	$50–70	$70–90	$90–110	$110–130
Number of workers.	14	109	301	278	156	73

Wage.....	$130–150	$150–170	$170–190	$190–210	$210–230	$230–250
Number of workers.	39	15	7	4	2	2

Find the quartile, mean, and standard deviations and also the decile points. Ans. $Q = 18.9$, $A.D. = 22.6$, $\sigma = 31.3$.

3. The problems at the end of Chapter IV furnish a large number of frequency distributions for which the measures of dispersion may be computed. The results will be of real interest in the case of problems 1, 2, 4, 6, and 7. For this purpose it is sometimes satisfactory to take wider intervals than those given in that chapter; the results are then less accurate but often not enough less to reduce their value for the purposes for which they would probably be used.

Digitized by Google

CHAPTER VIII

REGRESSION OR TREND LINES

HISTORICAL series rarely rise or fall with complete regularity. Sometimes it is evident on inspection that there is a general tendency for the later terms to be larger than the earlier terms, and the amount of change from term to term may be somewhat uniform. In other cases it seems doubtful whether there is any general tendency. It is therefore desirable to have definite methods for determining whether such a tendency exists, and if so, what the average amount of increase from term to term is. We shall discuss in this chapter the methods which have been found most valuable for these purposes.

Let us approach the problem by considering the various methods which might suggest themselves in determining the general tendency of the production of lumber in the United States. For the sake of ease in manipulation, let us assume that a preliminary examination of all available figures indicates that the true trend may be secured by analysis of the production figures for the years 1916 to 1922 inclusive, which are as follows in billions of board feet:

1916....39.8	1920....33.8
1917....35.8	1921....27.0
1918....31.9	1922....31.6
1919....34.6	

Method of selected values. One way which is popular among practical men, in spite of its crudity, is to determine the average decrease from year to year by taking two years at a time, preferably selecting years which seem typical. For instance, the average annual increase from 1916 to 1922 may be called $(31.6 - 39.8)/6 = -1.4$. Or we might say $(27.0 - 35.8)/4 = -2.2$, a result which is clearly unsatisfactory because it uses the abnormally low production of 1921. If we use 1917 and 1920, we find an average annual increase of $(33.8 - 35.8)/3 = -.67$. Similarly 1916 and 1920 give

Digitized by Google

−1.5, and 1917 and 1922 give −.8. The only way to make this method satisfactory logically would be to use all possible pairs of years and average the results, or else to use all but obviously exceptional pairs. But in the problems which actually arise, in which twenty or thirty years may be included, the number of possible pairs is so great that this method is not practicable. The only usable modification is selecting typical years carefully and using the result as a rough guide or check only.

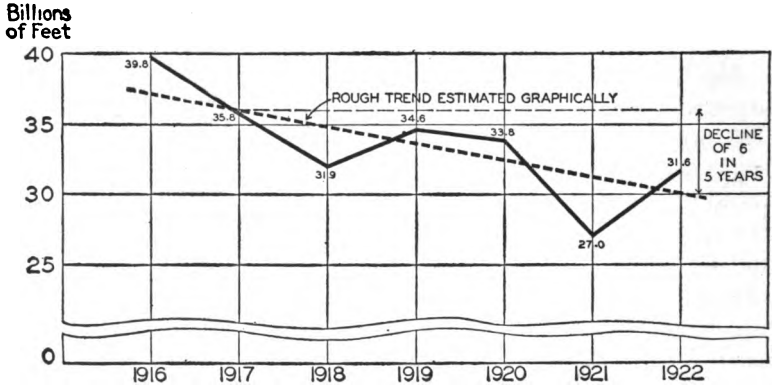


DIAGRAM XX. TREND OF LUMBER PRODUCTION IN THE UNITED STATES, 1916–1922

Graphic method. If the series is plotted, as in Diagram XX, it may be possible by inspection and the use of a straight edge or a black thread to locate graphically a line which seems to express the general trend of the plotted points. The slope of this line, that is to say, the ratio of vertical rise for any number of years to the corresponding number of years, gives average rise per year. This method is superior to the preceding, but is not capable of high refinement and gives different results with different workers. In our particular problem, it leads to an estimate of about −1.2 billions of board feet per year as the average slope. That is to say, on the average the annual production decreases by about 1,200,000,000 feet per year. This process is so rough, however, that we cannot feel confident that this summary of the facts is correct.

Digitized by Google

Method of adjustment of differences. In the case of series which are more regular than the one used as an example here, it is sometimes possible to smooth by adjustment of the differences between successive terms. This adjustment may be made by inspection or by algebraic formulas. The details will not be discussed here, as successful application of the method requires rather more mathematical training than is assumed here. The reader who is interested is referred to textbooks on the Theory of Finite Differences, to Chapter III in the *Handbook of Mathematical Statistics*, and to the references given at the end of that chapter.

Algebraic method. Another method, which is the one most generally used for careful work, is based upon algebraic formulas. The argument back of these formulas is in the simplest case the same as that of the graphic method discussed above, that the best trend to use is the straight line which is in some sense the best fit to the broken line graph of the given series. Expressing this argument in algebraic form, and using certain concepts from the previous chapter, on measures of dispersion, we derive formulas which serve as a basis for a numerical process which can be applied easily in any particular case.

The equation of a straight line can be expressed in the form

$$y = mx + b$$

where "m" is the slope of the line or the rate of change of y with respect to x, and "b" is the value of y when $x = 0$.[1] In order to get the particular straight line which will be the best fit to the given broken line, we must assign the proper values to these constants "m" and "b." As a first requirement, let us agree to choose them in such a way that the vertical distances from the line to the points representing the terms of the series shall add up algebraically to zero, or in other words, in such a way that the sum of the vertical distances from the line to the points above the line shall be equal to the sum of the distances from the line to the points below the line. This requirement would probably be adopted by almost any one who thought about the problem.

[1] Cf. *Introduction to Mathematical Analysis* by F. L. Griffin, chap. VIII.

Digitized by Google

Let us express this idea in symbolic form. The ordinate of the line corresponding to a time t_1 years after some selected base year is given by $mt_1 + b$. Denote by y_1 the actual lumber production in that same year. Then $y_1 - mt_1 - b$ is the excess of the actual ordinate over the ordinate of the line; for points below the line this expression will represent a negative number. The algebraic sum of such differences formed for all points should then be zero, to satisfy the requirement just proposed. Or in symbols,

$$\Sigma(y - mt - b) = 0.$$

If the number of points is N, this equation becomes

$$\Sigma y - m\Sigma t - Nb = 0.$$

But by definition of the arithmetic mean, $\Sigma y = N\bar{y}$ and $\Sigma t = N\bar{t}$, and hence

$$N\bar{y} - mN\bar{t} - Nb = 0, \text{ or } b = \bar{y} - m\bar{t}.$$

Substituting this value of b in the original equation, it takes the form

$$y = mt + \bar{y} - m\bar{t}, \text{ or } y - \bar{y} = m(t - \bar{t}).$$

Now $y - \bar{y}$ is the deviation of the ordinate from the mean ordinate, and $t - \bar{t}$ the number of years after the mean year. These quantities may well be denoted by the single letters Y and T, and the equation becomes

$$Y = mT. \qquad \text{(Formula VIII–1)}$$

The equation now contains only one constant, the slope, and the next step is to determine that, in other words, to pick out one of the family of lines through the "center" (the point on the ordinate for the mean year, at a height equal to the mean ordinate).

To determine m, we make use of the standard deviation, discussed in Chapter VII. We have just determined b by equating to zero the algebraic sum of the deviations of the given points from the straight line. A second requirement which might suggest itself would be to make as small as possible some measure of dispersion for these same deviations. As the mean or quartile deviations are not amenable to algebraic treatment, the standard deviation is actually used for this purpose. An argument for determining lines or curves by making the standard deviation as small as possible will be given in a

Digitized by Google

later chapter, in connection with the theory of probability. At present, it is merely proposed as a common-sense basis for selecting among the various possible lines.

In the notation finally adopted above, the deviation for the typical point is $Y - mT$, and the square of the standard deviation is therefore $\Sigma(Y - mT)^2$, or, expanding, $\Sigma Y^2 - 2m\Sigma YT + m^2\Sigma T^2$. We want to select the value of m which will make this expression as small as possible. It is shown,[1] in textbooks of algebra or calculus, that when a is positive, the quadratic function of x, $ax^2 + bx + c$, has its minimum when $x = -b/2a$. In our problem, then, the minimum value of the function of m occurs when $m = \Sigma YT/\Sigma T^2$. As $Y = y - \bar{y}$ and $T = t - \bar{t}$, in the original notation,

$$m = \frac{\Sigma(y - \bar{y})\,(t - \bar{t})}{\Sigma(t - \bar{t})^2}$$

This is sometimes called the *regression coefficient* of y on t, and the equation

$$y - \bar{y} = \frac{\Sigma(y - \bar{y})\,(t - \bar{t})}{\Sigma(t - \bar{t})^2}(t - \bar{t}) \text{ or } y - \bar{y} = m(t - \bar{t}) \text{ (Formula VIII–2)}$$

is called the *regression equation.* This terminology is more appropriate in connection with functional quantitative series than with historical series; in the latter case it is better to call m the *annual rate of increase* and to call the equation the *equation of trend.*

Before carrying through the arithmetic required by these formulas for the lumber production problem, it is desirable to prove a correction formula which enables us to use instead of the deviations about the true means, deviations about assumed means. Let Y_1' and T_1' be the distances from the assumed means to the first point, and d and c respectively the amounts by which the true means exceed the assumed means. Then

$$Y_1' = Y_1 + d, \; T_1' = T_1 + c.$$

Multiplying together, we have

$$Y_1'T_1' = Y_1T_1 + c\,Y_1 + dT_1 + cd.$$

[1] Cf. *Introduction to Mathematical Analysis* by F. L. Griffin, pp. 82 and 94, or *College Algebra* by H. B. Fine, pp. 307, 308.

Digitized by Google

Adding together the corresponding equations for all N points, we have

$$\Sigma Y'T' = \Sigma YT + c\Sigma Y + d\Sigma T + Ncd.$$

Now Y and T are deviations about the true means, and therefore $\Sigma Y = \Sigma T = 0$. Hence

$$\Sigma YT = \Sigma Y'T' - Ncd. \qquad \text{(Formula VIII–3)}$$

An important result of this formula is that if we use true mean for the time, c is zero, and therefore the correction is zero, even if we use the original y's for the Y's.

The computation can be carried out conveniently as indicated in the following table:

Year t	Diff. from Base Year $T = t - \bar{t}$	Production y	Product yT		Denominator T^2	Production by Formula $\bar{y} + mT$	Diff. Act.] − Formula $y-(\bar{y}+mT)$
1916....	−3	39.8	−119.4		9	37.8	2.0
1917....	−2	35.8	− 71.6		4	36.4	− .6
1918....	−1	31.9	− 31.9		1	34.9	−3.0
1919....	0	34.6	0		0	33.5	1.1
1920....	1	33.8		33.8	1	32.1	1.7
1921....	2	27.0		54.0	4	30.6	−3.6
1922....	3	31.6		94.8	9	29.2	2.4
Totals		234.5	−222.9	182.6	28		

$$\bar{y} = 234.5/7 = 33.5 \qquad m = \frac{\Sigma YT}{\Sigma T^2} = (-222.9 + 182.6)/28 = -1.43$$

Trend line is: $y = 33.5 - 1.43\,(t - 1919)$

It is to be noted that by using the true mean time, the correction to apply to the product was zero, even though the actual y's were used rather than deviations about the true mean of the y's. Particular attention should be given to the last two columns in the table, which of course cannot be filled in until m has been computed. The next to the last column shows the production according to the formula derived by this computation, and the last column shows how much these results differ from the actual production.

Digitized by Google

These differences should always be found and examined, both as a test of the accuracy of the arithmetic, and as an aid in deciding whether the trend is to be accepted as a significant summary of the facts. The differences should be small, and should not be consistently plus or minus for any considerable number of consecutive values. If these conditions are not satisfied, first check the arithmetic; if the arithmetic seems satisfactory, then try the methods described in the next section.

Methods to be used if straight line is a poor fit. If examination of the differences in the last column of the table, or inspection of the graph of the data, even before we try to find any formula, shows that a straight line will be a poor fit, the following methods are available:

(1) We may omit points which are clearly exceptional, and fit a line to the remainder.
(2) We may fit one straight line to a number of points at the beginning of the series, and another straight line to the remainder. This method is difficult to justify unless it is known underlying conditions have changed.
(3) We may take the logarithms of y and fit a straight line to them just as was done above to the y's themselves. We find as the equation in this case $\log y =$ (mean of the logarithms of the y's) $+ mx$, an equation which can also be put in the form $y = ae^{kx}$. We can test whether this equation will give a good fit by plotting the data on semi-logarithmic paper and noting how near the result is to a straight line. This equation means that the increase in the quantity in a year is a constant percentage of the value of the quantity at the beginning of the year. Such a law of increase seems to represent the facts very well for a good many series, so far as the past and immediate present is concerned, but usually leads to unbelievable results if projected more than a few years into the future.
(4) We may determine the constants in more complicated formulas by the same method as used above for the straight line.

Digitized by Google

The results are of course more complicated than in the case of the straight line. For that reason they are not used as frequently, but the increasing care with which various problems are being analyzed has already resulted in some use of this method, and may be expected to lead to further use. We shall illustrate the method by carrying through the algebra for the simplest formula for these purposes after the straight line and exponential curve, namely the parabola $y = ax^2 + bx + c$

Determination of constants in the quadratic formula $ax^2 + bx + c$. As in the case of the straight line, the first step is to determine some constant so that the algebraic sum of the differences between the observed and the computed values shall be equal to zero. We have, then,

$$\Sigma(y - ax^2 - bx - c) = 0 \quad \text{or} \quad \Sigma y - a\Sigma x^2 - b\Sigma x - Nc = 0.$$

But we know that

$$\Sigma x = N\bar{x}, \Sigma y = N\bar{y}, \quad \text{and} \quad \Sigma x^2 = N\sigma_x^2 = N(\sigma_x^2 + \bar{x}^2),$$

and hence

$$N\bar{y} - aN(\sigma_x^2 + \bar{x}^2) - bN\bar{x} - Nc = 0 \quad \text{or} \quad c = \bar{y} - b\bar{x} - a(\sigma_x^2 + \bar{x}^2).$$

Substituting this value of c in the formula, we have:

$$y = ax^2 + bx + \bar{y} - b\bar{x} - a(\sigma_x^2 + \bar{x}^2), \text{ or } y - \bar{y} = b(x - \bar{x}) + a(x^2 - \sigma_x^2 - \bar{x}^2).$$

If, as in the case of the straight line, we now write $y - \bar{y} = Y, x - \bar{x} = X$, we find

$$Y = bX + a(X^2 + 2X\bar{x} + \bar{x}^2 - \sigma_x^2 - \bar{x}^2), \text{ or} \quad \text{(Formula}$$
$$Y = (b + 2a\bar{x})X + a(X^2 - \sigma_x^2) = BX + a(X^2 - \sigma_x^2) \quad \text{VIII–4)}$$

We have thus found a formula with one less constant than the original formula, just as we found Formula VIII–1 for the straight line. The next step is, as before, to determine the remaining constants so as to minimize the sum of the squares of the differences between the observed values and the values computed from this formula, that is to say, of

$$\Sigma[Y - BX - a(X^2 - \sigma_x^2)]^2, \text{ which we shall call } K.$$

Expanding, we have

Digitized by Google

$$K = \Sigma Y^2 + B^2\Sigma X^2 + a^2\Sigma(X^2 - \sigma_x^2)^2 - 2B\Sigma XY - 2a\Sigma X^2Y + 2a\sigma_x^2\Sigma Y + 2aB\Sigma X^3 - 2aB\sigma_x^2\Sigma X.$$

Now $\Sigma X = \Sigma Y = 0$, as X and Y are deviations about the mean, and also

$$\Sigma X^2 = N\sigma_x^2, \ \Sigma Y^2 = N\sigma_y^2.$$

We have, then,

$$\Sigma(X^2 - \sigma_x^2)^2 = \Sigma(X^4 - 2\sigma_x^2X^2 + \sigma_x^4)$$
$$= \Sigma X^4 - 2\sigma_x^2N\sigma_x^2 + N\sigma_x^4 = \Sigma X^4 - \sigma_x^2\Sigma X^2$$
$$K = B^2\Sigma X^2 + 2aB\Sigma X^3 + a^2(\Sigma X^4 - \sigma_x^2\Sigma X^2) - 2B\Sigma XY - 2a\Sigma X^2Y + \Sigma Y^2.$$

We desire to choose such values of a and B that K shall be as small as possible. Now it is shown in advanced calculus [1] that the expression

$$lx^2 + 2txy + my^2 + 2ux + 2vy + s$$

has its minimum value, if l and m are both positive and lm is greater than t^2, when

$$lx + ty + u = 0$$
$$tx + my + v = 0.$$

In this case, then, K will have its minimum value when

$$B\Sigma X^2 + a\Sigma X^3 - \Sigma XY = 0 \qquad \text{(Formulas VIII–5)}$$
$$B\Sigma X^3 + a(\Sigma X^4 - \sigma_x^2\Sigma X^2) - \Sigma X^2Y = 0,$$

if, as will be shown later, $\Sigma X^4 - \sigma_x^2\Sigma X^2$ is positive and

$$(\Sigma X^2)\ (\Sigma X^4 - \sigma_x^2\Sigma X^2) \text{ is greater than } (\Sigma X^3)^2.$$

As all the summations can be found from the given data, formulas VIII–5 enable us to determine a and B by the solution of two simultaneous linear equations.

These equations as they stand are general and apply to any historical series. They may be materially simplified if, as is usually the case, the successive intervals of time are equal throughout and no years (or quarters, or months) are omitted from consideration. In such cases there is symmetry with respect to x (the time) and

[1] See *Elements of the Differential and Integral Calculus*, by W. A. Granville, pp. 243–46.

Digitized by Google

therefore the sum of the cubes of the positive deviations from the mean time is numerically equal to the sum of the cubes of the negative deviations, and therefore the algebraic sum of all such deviations, i.e., ΣX^3, is equal to zero. From formulas VIII–5 we then find at once:

$$B = \Sigma XY / \Sigma X^2$$

and $$a = \Sigma X^2 Y / (\Sigma X^4 - \sigma_x^2 \Sigma X^2) \qquad \text{(Formulas VIII–6)}$$

It should be noted that this value of B is exactly the same as the coefficient of X in the equation of the straight line fitted to the same data.

As we shall show in the next paragraph, the denominators of B and a can be read from a table, or computed from simple formulas based on the number of terms in the series. The only real addition to the computation shown on page 160, in order to adapt it to the determination of the parabola, is the addition of a column in which we multiply each "Xy" again by "X." The sum of the entries in this column gives $\Sigma X^2 y$. As $y = \bar{y} + Y$, $\Sigma X^2 y = \bar{y}\Sigma X^2 + \Sigma X^2 Y$, or $\Sigma X^2 Y = \Sigma X^2 y - \bar{y}\Sigma X^2$, and hence the numerator of a can be found.

Formulas for the denominators in formulas VIII–6. Tables in the appendix give the value of the denominators of B and a for N (number of terms) = 1 to $N = 100$. These tables are based on the following formulas, which will be proved below:

$$\Sigma X^2 = N(N^2 - 1)/12, \quad \Sigma X^4 - \sigma^2 \Sigma X^2 = N(N^2 - 1)(N^2 - 4)/180 \qquad \text{(Formulas VIII–7)}$$

As the proof of these formulas is merely algebraic, without any statistical implications, many readers will be content to accept the formulas without following through the details of the proof, or even to use the table without paying attention to the formulas. The proof is included, however, for the use of those who desire to understand every step as completely as possible and to be able to make such extensions and modifications as unusual problems may require.

The following formulas are proved in some algebras, or may be proved by mathematical induction:

If N is any positive integer,

Digitized by Google

$1 + 2 + 3 + \ldots + N$ [denoted by ΣN] $= N(N+1)/2$
$1^2 + 2^2 + 3^2 + \ldots + N^2$ [denoted by ΣN^2] $= N(2N+1)(N+1)/6$
$1^3 + 2^3 + 3^3 + \ldots + N^3$ [denoted by ΣN^3] $= N^2(N+1)^2/4$
$1^4 + 2^4 + 3^4 + \ldots + N^4$ [denoted by ΣN^4] $=$

$$\frac{N(N+1)(6N^3+9N^2+N-1)}{30}$$

Now for a time series with equal intervals and no breaks, we may choose our origin so that x takes in succession the values 1, 2, 3, . . . up to N. Then the mean x, $\bar{x}$, $= \Sigma N/N = (N+1)/2$, by the first formula above, and

$$\Sigma X^2 = \Sigma(x-\bar{x})^2 = \Sigma(x^2 - 2x\bar{x} + \bar{x}^2) = \Sigma N^2 - 2\bar{x}\Sigma x + N\bar{x}^2$$

$$= N(N+1)(2N+1)/6 - (N+1)^2/4 = \frac{N(N+1)}{12}(4N+2-3N-3)$$

$$= N(N^2-1)/12.$$

We have also $\sigma_x^2 = \Sigma X^2/N = (N^2-1)/12$.

Similarly,

$$\Sigma X^4 = \Sigma(x-\bar{x})^4 = \Sigma(x^4 - 4x^3\bar{x} + 6x^2\bar{x}^2 - 4x\bar{x}^3 + \bar{x}^4)$$

$$= \Sigma N^4 - 4\frac{(N+1)}{2}\Sigma N^3 + 6\frac{(N+1)^2}{4}\Sigma N^2 - 4\frac{(N+1)^3}{8}\Sigma N$$

$$+ N\frac{(N+1)^4}{16}$$

$$= \frac{N(N+1)(6N^3+9N^2+N-1)}{30} - \frac{N^2(N+1)^3}{2}$$

$$+ \frac{N(N+1)^3(2N+1)}{4} - 3N(N+1)^4/16$$

$$= \frac{N(N+1)(6N^3+9N^2+N-1)}{30}$$

$$+ N(N+1)^3[8N+4-8N-3N-3]/16$$

After a little simplification, we have:

$$\Sigma X^4 = N(N+1)(N-1)(3N^2-7)/240.$$

The denominator of a in formulas VIII–5 is $\Sigma X^4 - \sigma_x^2 \Sigma X^2$, which may now be simplified.

Digitized by Google

$\Sigma X^4 - \sigma^2 \Sigma X^2$
$= N(N+1)(N-1)(3N^2-7)/240 - (N^2-1)N(N^2-1)/144$
$= N(N^2-1)[(3N^2-7)3 - 5(N^2-1)]/720$
$= N(N^2-1)(N^2-4)/180.$

Applications of formulas VIII–4, VIII–6, and VIII–7. The use of the preceding formulas will be clearer to the reader if he will follow through the steps of the application in two problems. The first is numerically very simple, for purposes of illustration merely; the second is the determination of the best parabolic trend for American automobile production, on the basis of the years 1912 to 1923 inclusive.

Let us find the parabola of closest fit to a quantity whose value in five successive years was 10, 12, 11, 13, 15. The computation is:

Year x	Deviation from Mid-Year (3) X	Quantity y	Product Xy	Second Product X^2y	y by Formula: Str. line $\bar{y} + BX$	y by Formula: Parab. y_p
1.....	−2	10	−20	40	10	10.4
2.....	−1	12	−12	12	11.1	10.9
3.....	0	11	0	0	12.2	11.8
4.....	1	13	13	13	13.3	13.1
5.....	2	15	30	60	14.4	14.8
Totals...		61	11	125		

We have $\Sigma y = 61 = 5\bar{y}$, $\bar{y} = 12.2$, $\Sigma Xy = 11$, $\Sigma X^2y = 125$.

From the tables, the denominator of a is 14 and of B is 10, when $N = 5$.

We have, then, σ_x^2 = denominator of B divided by 5 = 2.

$B = \Sigma Xy/10 = 11/10 = 1.1$
$a = (\Sigma X^2y - \bar{y}\Sigma X^2)/14 = (125 - 12.2 \times 10)/14 = 3/14 = .214.$

The equation of the straight line is therefore $y = \bar{y} + 1.1X$, and of the parabola,

$y = \bar{y} + BX + a(X^2 - \sigma_x^2)$ or $y = 12.2 + 1.1X + .214(X^2 - 2)$ or

Digitized by Google

$$y = 11.8 + 1.1\,X + .214\,X^2 = y_p, \text{ say.}$$

As $X = x - 3$, this equation may be reduced to the form:

$$y = 10.4 - .184x + .214x^2$$

Substituting in succession the values, 1, 2, 3, 4, 5 for the year x in this equation, or $-2 \ldots + 2$ for X in y_p, we find for successive values of the quantity 10.4, 10.9, 11.8, 13.1, 14.8, which it will be noted, are reasonably close to the given values. The linear regression equation gives 10, 11.1, 12.2, 13.3, 14.4; it will be found that the sum of the squares of the differences of actual and computed values is 2.7 for the linear formula and 2.05 for the quadratic, showing that the latter is distinctly better by the least squares criterion. Whether this increased closeness of fit is valuable enough to justify the additional labor of computation, is of course to be decided by a consideration of the purposes for which the result will be used.

The computation for the trend of automobile production is as follows:

Year	Deviation from 1917½ (X)	Production of Automobiles in Thousands (y)	Product (Xy) Times 2	Second Product (X^2y) Times 4
1912....	−11/2	378	−4158	45738
1913....	−9/2	485	−4365	39285
1914....	−7/2	574	−4018	28126
1915....	−5/2	893	−4465	22325
1916....	−3/2	1584	−4752	14256
1917....	−1/2	1869	−1869	1869
1918....	1/2	1154	1154	1154
1919....	3/2	1876	5628	16884
1920....	5/2	2205	11025	55125
1921....	7/2	1669	11683	81781
1922....	9/2	2576	23184	208656
1923....	11/2	4013	44143	485573
Totals.....		19276	−23627 + 96817 2|+73190 36595	4|1000772 250193

Digitized by Google

That is to say,

$$\Sigma y = 19{,}276 = 12\bar{y},\ \bar{y} = 1{,}606$$
$$\Sigma Xy = 36595,\ \Sigma X^2y = 250{,}193$$

From the tables, for $N = 12$, the denominator of a is $\frac{4004}{3}$ and of B, 143.

We then have: $\sigma_x^2 =$ denom. of B divided by $12 = \frac{143}{12} = 11.917$.

$$\Sigma X^2Y = \Sigma X^2y - \sigma_x^2 \Sigma y = 250{,}193 - 11.917 \times 19{,}276 = 20{,}481$$

$$B = \frac{\Sigma Xy}{143} = \frac{36{,}595}{143} = 255.9$$

$$a = \frac{3\,\Sigma X^2Y}{4004} = \frac{61{,}443}{4004} = 15.345$$

The equation therefore is:

$$\text{Production} = 1606 + 255.9\,(\text{Year} - 1917.5)$$
$$+ 15.345\,[(\text{Year} - 1917.5)^2 - 11.917]$$

which, by writing Year $-$ 1917.5 $=$ (Year $-$ 1918 $+ \frac{1}{2}$) and simplifying, we may reduce to

$$\text{Production} = 1555 + 271.2\,(\text{Year} - 1918) + 15.345\,(\text{Year} - 1918)^2.$$

The validity of this equation should be checked by finding the production for each year according to the equation and comparing it with the actual production. The difference (Actual–formula) should be fairly small, and should not be of the same sign for many consecutive values. The following table shows this comparison, for the formula just found, and also for the formula $y = ae^{cx}$, in which the constants are determined by taking logarithms and then proceeding exactly as was done in finding linear regression equations, using log y instead of the y of the formula.

Both of these curves are defective in giving too large values for the first years with errors which are a large percentage of the actual figures. It should be noted that the positive differences add up to 1891 and the negative to -1890 in the case of the parabolic curve, while in the case of the exponential these totals are 2081 and 1772

Digitized by Google

respectively. The algebraic sum of these differences (i.e., the errors) is therefore practically zero, in the former case, as it should be because this identity was assumed in deriving the formula used. The

Year	Actual Production	Parabolic Formula		Exponential Formula	
		Production	Difference Actual–Formula	Production	Difference Actual–Formula
1912...	378	480	−102	468	− 90
1913...	485	583	− 98	564	− 79
1914...	574	716	−142	678	−104
1915...	893	880	13	817	76
1916...	1584	1074	510	983	601
1917...	1869	1299	570	1183	686
1918...	1154	1555	−401	1424	−270
1919...	1876	1842	34	1714	162
1920...	2205	2159	46	2063	142
1921...	1669	2507	−838	2484	−815
1922...	2576	2885	−309	2990	−414
1923...	4013	3295	718	3599	414

total deviation is 3781 for the parabola, as compared with 3853 for the exponential curve but the sum of the squares of the errors is less for the exponential curve.

Use of trend lines in forecasting the future. We have considered trend lines up to this point as a device for eliminating the effect of temporary and accidental variations, in order to determine whether there is any general tendency for later terms to be larger or smaller than earlier terms, and if so, what the average amount of change from term to term is. One major use for trend lines, however, has not been considered explicitly. That use is as an aid in forecasting the future.

The student who has become accustomed to the exact methods of pure mathematics is ordinarily reluctant to extend a trend line into the future and attach any importance to the results which it seems to predict. This reluctance is in the main justified. But if he will

Digitized by Google

approach the problem from a somewhat different point of view, he will see more justification for using a carefully checked trend line as a guide in estimating future conditions. He should place himself in the position of the business man or statesman who must find all light he can to guide him. Consider some of the problems a business man is called on to answer every day, often with possibilities of great financial loss if he decides wrongly. For instance, shall he buy his raw material to-day or wait — in other words, are prices going to rise or fall? Shall he manufacture ahead of orders, or lay off all workmen not absolutely required — in other words, will the future bring more or fewer orders than at present? Shall he grant credit to Mr. A, whose financial condition depends on a continuance of good general business conditions?

All of these questions are very difficult to deal with, and in answering them the business men or their advisers will need all the help they can get from every source. If they can know that the natural continuance of growth as in the past would lead to sales of about so much, they have a definite basis to work on. The figures so found will probably be modified to allow for new conditions, but the final result will be more reliable than if mere unaided "judgment" were used to gauge the effect of natural growth and new conditions all at once.

In order to get a trend line which will be satisfactory when tested from this point of view, it is necessary to examine the results of any formal mathematical process with great care. We should consider whether the position of the trend with respect to the actual figures, especially in the more recent years, is in accord with a reasonable interpretation of the facts in the light of all available information. By choosing different periods to work with in determining the constants, and using different formulas and methods of fitting, we can always secure a number of different possible trends. We should then select from among these after careful consideration of all factors. It is usually helpful to plot the actual values and all the trends on the same field, extending the latter some years into the future. It is unfortunately impossible to give an example here, with full explanation of the reasons for choosing one trend rather than another, as some of the reasons will be based on a detailed

Digitized by Google

and intimate knowledge of the particular problem. It is possible merely to point out that careful consideration is necessary, and to warn against accepting a trend merely because the mathematical processes have been carried out correctly.

Parabolic and exponential trends, when carried into the future, often rise above what we know from other considerations to be reasonable limits. A device which has sometimes been acceptable in such cases is to extend into the future along the tangent to the curve at some recent point. That is to say, we estimate that while the amount of increase per year has been increasing up to the present time, future increases will be at the present rate.

Some readers may think that mathematical processes are of no value, if they are to be modified and interpreted as advocated in the last two paragraphs. The main advantage is that after a trend has been adopted for future use, current developments can be interpreted on a consistent basis without reconsideration of the basic factors. Of course, the whole thing should be reconsidered if it is known that conditions have changed materially, or if the results seem unreasonable. But usually such reconsideration is not required oftener than once in a year.

Trend lines for functional quantitative series. The formulas developed in this chapter are applicable almost without change to functional quantitative series. The numerical work is often more complicated, however, because a number of values of y correspond to a single value of x, or else the values of x are not equi-distant. An example will be worked through to show how to deal with these additional complications.

During a period of sixty years, the average number of bushels of corn per acre in Ohio varied with the July rainfall as follows:

Rainfall.....	1–2	2–3	3–4	4–5	5–6	6–7	7–8	8–9
No. of years with above rainfall.....	2	12	16	16	10	3	0	1
Average yield (bushels per acre).....	30.0	31.7	34.4	34.4	38.0	37.5		42.5

It is clear that the average yield per acre is larger when the rain-

Digitized by Google

fall in July is larger. But the increase is somewhat irregular. This is not unexpected, in view of the many circumstances which are not considered in the figures. Amount of rainfall in other months, amount of sunshine, labor conditions, price of corn, variation in acreage, and many other variables, affect the yield, in addition to July rainfall. We have a quantity which is in fact a function of a large number of variables, and we try to separate out one of them, in the hope that the effects due to the other variables will not, on the average, have much influence on the results. The validity of this hope will be analyzed in later chapters, in connection with the general discussion of probability.

If we ignore, then, for the time being the variables other than July rainfall and yield in bushels per acre, we may summarize the general trend of the table by a statement of the average yield and a statement that on the whole and on the average an increase of one inch in July rainfall is accompanied by an increase of about so much in the yield per acre. The determination numerically of *how* much is exactly the same problem as we have just solved for historical series, except that instead of time as the independent variable we use rainfall. In the historical series, there is only one value of y for each value of the time, but in the functional quantitative series there are a number of years in which the rainfall had the value 1–2, 2–3, etc. As the summations in the formulas are to apply to *all* instances, it is best to write them with an "f" (frequency) as a reminder of the fact that each term is to be taken a number of times. The formulas then become:

$$\text{Mean rainfall} = \bar{x} = \Sigma fx/N, \quad \text{Mean yield} = \bar{y} = \Sigma fy/N,$$

$$m = \Sigma fXY/\Sigma fX^2 = \Sigma f(x-\bar{x})\ (y-\bar{y})/\Sigma f(x-\bar{x})^2$$

$$\Sigma fXY = \Sigma fX'Y' - Ncd, \quad \Sigma fX^2 = \Sigma fX'^2 - Nc^2,$$

where X' and Y' are deviations about assumed means for x and y respectively, and X and Y are deviations about the true means.

The process of computing the constants in the regression equation for the yield-rainfall example is shown in the following table:

Digitized by Google

Rainfall x	Yield y	Frequency f	Deviations fr. Assumed Means $x-4.5$ X'	$y-35$ Y'	Freq. × Dev. fX'	Freq. × Dev. fY'	Product $fX'Y'$	Denom. fX'^2	Yield by Form. $\bar{y}+mX$
1.5.....	30.	2	−3	−5.	− 6	−10.	30.	18	30.2
2.5.....	31.7	12	−2	−3.3	−24	−39.6	79.2	48	31.9
3.5.....	34.4	16	−1	−0.6	−16	− 9.6	9.6	16	33.6
4.5.....	34.4	16	0	−0.6	0	− 9.6	0.	0	35.3
5.5.....	38.0	10	1	3.0	10	30.	30.	10	37.0
6.5.....	37.5	3	2	2.5	6	7.5	15.	12	38.7
7.5.....		0							40.4
8.5.....	42.5	1	4	7.5	4	7.5	30.	16	42.1
Totals.....		60			−26	−23.8	193.8	120	

From these totals we find:

$c = \Sigma fX'/N = -26/60 = -.43$, and hence $\bar{x} = 4.5 - .43 = 4.07$.

$d = \Sigma fY'/N = -23.8/60 = -.4$, and hence $\bar{y} = 35 - .4 = 34.6$.

$\Sigma fX^2 = \Sigma fX'^2 - Nc \times c = 120 - (26)(.43) = 108.8$.

$\Sigma fXY = \Sigma fX'Y' - Ncd = 193.8 - (-26)(.4) = 183.4$.

$m = \Sigma fXY/\Sigma fX^2 = 183.4/100.8 = 1.69$.

Regression equation: $y = 34.6 + 1.69\,(x - 4.07)$.

Or, in words, the average yield is 34.6, corresponding to a rainfall in July of 4.07 inches; a change in rainfall of one inch leads to a change of 1.69 bushels in the yield per acre.

In view of the small number of years on which these relationships are based, it is desirable to round off these figures still more before quoting them as significant, and to say that the average yield is 35 bushels per acre, corresponding to a rainfall of four inches, and that a change of one inch in rainfall leads to a change of 1.7 bushels in the yield per acre.

The last column of the table is computed *after* the rest of the work has been done, and serves as a check on its accuracy and also on the closeness of fit of the formula assumed, in this case a straight line.

Digitized by Google

General remarks on selection of an equation. Before detailed computations to determine the constants in a particular formula are undertaken, careful reconnoitering should be done to make reasonably sure that a satisfactory formula has been selected. Familiarity with various curves which have proved useful in similar problems is desirable, and also consultation of standard works on curve fitting, for example, T. R. Running's *Empirical Formulas* and J. Lipka's *Graphical and Mechanical Computation.*

Plotting the given data on various kinds of graph paper should be part of the routine in formula selection. If the graph on ordinary paper is well represented by a straight line, well and good; the search is over. But usually this will not be conclusive without at least trying out the graph on semi-logarithmic paper, plotting the first differences, or thinking of more complicated curves. As noted above, the exponential curve is indicated if the semi-logarithmic graph is straight. If the graph of the first differences is approximately straight, the ordinary parabola is indicated.

A curve which the author has found useful in many business series is the curve whose equation is:

$$y = \frac{a}{10^{bx} + c}$$

This curve is used in physical chemistry when the rate at which a reaction takes place is proportional both to the amount of the original agent and the amount of transformed material. From this property it is known as the *autocatalytic curve.* In statistical work, it was originally called the logistic curve, but more recently the Pearl-Reed curve, because of its use by Professors Pearl and Reed of Johns Hopkins in population problems. In combination with a straight line, it can be used to advantage for a series which starts small, increases rapidly, at almost an exponential rate, for a while, and then reverts to a moderate uniform rate of increase.

Summary of Chapter VIII. The general tendency or trend of historical series may be estimated by the following methods:

(1) Method of selected points — simple, but rough.

Digitized by Google

(2) Graphic method — easy, but gives different results in the hands of different persons.

(3) Method of adjustment of differences — applies only to rather regular series, and requires mathematical training.

(4) Algebraic method — after formulas have been derived, simple to use and definite in results.

Formulas are derived for fitting (1) a straight line, and (2) a parabola to a given series. If the logarithms of the terms are used rather than the terms themselves, these formulas lead to exponential curves.

When trend lines are projected into the future, they should be checked against all available information before any reliance is placed on them. If used with care, however, they furnish a valuable basis for predictions.

The formulas developed for historical series may be used with functional quantitative series with certain minor modifications.

EXAMPLES

1. The average weight of women 5 ft. 7 in. tall, was found to vary in a typical group of 115 as indicated in the following table:

Age	Number of Women	Average Weight
15–19	4	134
20–24	20	137
25–29	28	140
30–34	24	144
35–39	19	148
40–44	10	151
45–49	6	157
50–54	3	158
55–59	1	157

Find the equation of the straight line which best expresses weight in terms of age for this group.

2. The population of Oregon was as follows at the dates indicated. Find the best linear and parabolic trend lines, by the methods explained in this chapter.

Year.....	1880	1890	1900	1910	1920
Population.....	175,000	318,000	414,000	673,000	783,000

3. The average earnings of Medical graduates of Harvard University, Class of 1907, were as follows in their early years of practice:

Digitized by Google

Year of practice.....	1	2	3	4	5	6	7
Average earnings.....	$360	$600	$1350	$1960	$2350	$3200	$3550

Find the straight line which best expresses the relation of average earnings to year of practice.

4. Find the parabola of closest fit for the following series of the imports of raw silk into the United States.

1898	12,088,000 lbs.	1907	18,744,000 lbs.	1916	41,925,000 lbs.
1899	11,250,000	1908	16,662,000	1917	40,351,000
1900	13,074,000	1909	25,188,000	1918	43,681,000
1901	10,406,000	1910	23,457,000	1919	55,522,000
1902	14,235,000	1911	26,666,000	1920	39,660,000
1903	15,271,000	1912	26,585,000	1921	52,332,000
1904	16,723,000	1913	32,102,000	1922	58,467,000
1905	22,357,000	1914	34,546,000	1923	46,577,000
1906	17,352,000	1915	31,053,000		

Answer: $y = 27{,}116{,}000 + 1{,}855{,}500\,x + 44{,}500\,x^2$, where $x = 0$ in 1911.

Digitized by Google

CHAPTER IX

ANALYSIS OF TIME SERIES

One of the main jobs of the business statistician is the analysis of time series. In studying the record of past sales, for instance, he needs to have a definite measure of the extent to which the sales of a particular month or year were above or below what might be considered "normal." As he considers the current record of these quantities, it is even more necessary to have a standard with which to make comparisons, in order to say whether these records should be considered high or low.

In the preceding chapter, we discussed one of the important steps in the analysis of time series, the determination of the trend over a period of years. By taking the ratio of the actual quantity in a particular year to the ordinate of trend for that year, we obtain a series of numbers which are freed from the effect of trend. We may say, then, that we have *eliminated* the secular trend, and reduced the annual figures to a comparable basis.

In order to complete the analysis of *monthly* figures, however, it is necessary to consider the influence of another factor which often obscures the true condition of affairs even more than does the secular trend. This factor is the season of the year. For example, if we are analyzing the sales of anthracite coal, it is obvious that we cannot say whether a particular month's record is high or low until we know whether it was May, November, or February. The only true basis for judgment is a comparison of the current November record with previous Novembers. The practical business man recognizes this principle when he compares the current November with the latest previous November. His method, to be sure, is inadequate in that he has to modify his judgment in accordance with his recollection of peculiar conditions which may have affected the previous November. If we modify his method by making the standard of comparison the average of three or seven previous Novembers rather than the latest only, we have a method

Digitized by Google

that is more justifiable, although the effect of trend would interfere with sound judgment of the results.

Another method which has been suggested is to regard the November figures as a time series and determine its trend. The ordinates of trend so determined for November would not ordinarily, however, be comparable with similarly determined ordinates for October and December, nor would this trend indicate the same rate of increase per year as is given by the trend based on the annual figures. In fact the trend based on any series of monthly figures will be relatively unstable because of the greater variability of monthly as compared with annual figures. In view of these disadvantages, and the labor of computing thirteen trends and securing some sort of coördination among them, this method is not to be recommended and has seldom been used.

The method of dealing with seasonal variation which has usually been adopted is to determine *seasonal factors*, which express what percentage each month is of the average month. If the actual figures for any month are divided by the appropriate seasonal factor, the quotients show what the figures would be with seasonal variation eliminated.

Three different aspects of the problem of seasonal variation. As we see above, seasonal factors are obtained and used in order to obtain monthly figures with the effect of seasonal variation eliminated. They are also used for two somewhat different purposes, which should be clearly differentiated, as the choice of method to be used in determining the factors depends to some extent on the purpose for which they are to be used.

The second purpose is to exhibit the facts in regard to seasonal variation, as something significant in itself apart from the adjustment of series. It is of importance, for instance, to know that July and August in New York City are about 20 degrees warmer, and January and February about 20 degrees colder, than the average for the year. These facts are valuable even if we have no series to adjust or no desire to eliminate the seasonal variation. For this purpose it may be better to express the measures of seasonality as deviations rather than as factors.

Digitized by Google

Another aspect of the problem of seasonal variation is the estimation of the expected figure for the year on the basis of actual results for the first several months. As business men put it, what is the annual rate for the year? From the logical standpoint, the way to determine annual rates is to compute what percentage the sales (or other quantity) during the first four months (say) have been of the sales for the year, during each of the past 5, 10, or 20 years, and select from these per cents some typical value — median or mean. The number of years used should be as large as possible without including years in which there is reason to believe that conditions differ substantially from present conditions.

The annual rate factors derived by this method, however, will be influenced by the fact that in a growing series the later months of the year will be larger than the earlier, wholly aside from any seasonal effect, but merely because of the upward trend. They will therefore differ somewhat from factors derived so as to be independent of trend, which are what are desired for adjusting series or exhibiting seasonality. If factors determined for the latter purposes are used to give annual rate factors, they will show what the annual figures may be expected to be, if we disregard the probable increases in later months due to normal growth. In many cases, such estimates as result from this method, which are lower than may really be expected, are safer to plan on than the results which would follow from the true annual rate factors.

In our discussion, then, of methods of finding seasonal factors, we shall emphasize most the methods primarily devised for the first two purposes, as the results of those methods are as a matter of practice usually used for all purposes.

Summary of methods used to determine seasonal factors. While it is often possible to determine a satisfactory trend for a series and eliminate the trend before concerning oneself with seasonal variation, there are other cases in which it is inconvenient or impossible to do so. It is therefore desirable to have methods which can be applied directly to the series without removing the trend, as well as methods which assume that the trend is negligible or already removed, and methods which combine elimination of a

Digitized by Google

trend with determination of seasonal factors. It is also desirable to have methods as simple as possible, even at the expense of extreme accuracy, and also refined methods for cases in which considerable accuracy is both possible and requisite.

The various methods which may be used, then, are classified as follows:

A. Simple methods, applicable without modification if the series has no appreciable trend.
   1. Method of monthly means.
   2. Method of monthly per cent of annual.

B. Link-relative methods, applicable regardless of the existence of a trend.
   1. The geometric-mean L.R. method.
   2. The median or arithmetic-mean L.R. method.

C. Methods involving the determination of a trend.
   1. The method of typical ratio to twelve-months moving average.
   2. The method of typical ratio to trend expressed by an equation.
   3. Method of typical ratio to a free-hand curve.
   4. Method of monthly per cent of annual, modified to eliminate effect of trend.

It should be noted that the twelve-months moving average and free-hand curve, in C 1 and C 3 respectively, are not pure trends, but trends modified by the broad swings of the series — in business statistics often called the cyclical variations.

We shall next discuss in detail each of these methods, and work through the arithmetic in actual cases by several of the methods.

A 1. The method of monthly means. The procedure is what most investigators would probably try first. Take the arithmetic mean of all figures for January as typical of January, and similarly

Digitized by Google

for each of the other months. If seasonal factors are desired, divide the mean for each month by the mean for the year.

For purposes of illustration, let us analyze the seasonal variation of the temperature at New York City. The figures for seven years are shown below. For an accurate determination, of course, a longer period should be used.

MONTHLY AND ANNUAL MEAN TEMPERATURES AT NEW YORK CITY

YEAR	JAN.	FEB.	MAR.	APR.	MAY	JUNE	JULY	AUG.	SEPT.	OCT.	NOV.	DEC.	ANNUAL
1917....	32	28	39	47	53	68	74	75	63	52	41	25	50
1918....	22	30	41	50	64	66	73	75	63	59	46	39	52
1919....	35	35	42	49	61	70	74	70	66	58	44	30	52
1920....	24	29	41	48	58	68	72	73	67	60	44	38	52
1921....	34	35	48	55	60	70	76	71	71	56	45	34	55
1922....	29	34	41	51	64	71	73	71	67	58	45	34	53
1923....	31	27	37	49	59	72	72	71	67	56	45	42	52
Av.....	30	31	41	50	60	69	73	72	66	57	44	35	52
Median.	31	30	41	49	60	70	73	71	67	58	45	34	52

It will be noted that for no month does the median differ by more than a degree from the average, which is therefore a good typical figure.

We may now find the seasonal index of January by dividing the mean for that month (30) by the mean for the years (52). We thus obtain .58 or 58 per cent. For the other months the values may similarly be easily found to be Feb. 62, March 79, April 96, May 115, June 133, July 140, Aug. 138, Sept. 127, Oct. 110, Nov. 85, Dec. 67. These indices add up to 1210, for which excess over 1200 adjustment may be made roughly by decreasing all indices except the two smallest by one unit. If really accurate values are desired, the tenths should be retained in the annual temperatures and in the averages. The indices found are simpler, and accurate enough to show what is meant by seasonal indices. Let us consider them briefly:

Jan.....58	April..... 95	July.....139	Oct.....109
Feb.....62	May.....114	Aug.....137	Nov..... 84
March.....78	June.....132	Sept.....126	Dec..... 66

Digitized by Google

Inspection of these indices gives in one form a picture of the change of temperature with the seasons. Another way of seeing the same thing is to look at the monthly averages themselves and regard them as standards with which to compare the actual temperatures in any given year. We may note, for instance, that 1918 and 1920 had unusually cold Januaries, and that September, 1921, was abnormally hot.

In the application of this process to business series, it is often appropriate to state these relations more definitely. We may:

(1) Divide the actual figures for any month by the seasonal index for that month, thus obtaining figures from which the effect of seasonal influences have been removed.

(2) Divide the actual figure for any month by the monthly mean for that month, thus obtaining a "per cent of standard" or "ratio to normal."

As these adjustments are not of great significance in this series of temperatures, in which the results would be different according to the temperature-scale chosen, the complete working-out of this method will be postponed to a later example.

A 2. The method of monthly per cent of annual. A second simple method of determining seasonal variation is to find for each year what per cent of the total for the year the figure for each month is, and then select from the percentages for January in different years a typical value, and similarly for other months. The "typical value" may be the arithmetic mean, the median, the average of the middle three or four (sometimes called the "modified median"), or any other value which on inspection seems representative. In general the author recommends the modified median, although the arithmetic mean has the advantage of yielding indices which add up to 1200 without adjustment.

This method will be applied to the sales of a certain manufacturing company during the years 1919 to 1924 inclusive, and the results compared with the results of other methods to be described later. The record of sales by months was as follows (in thousands of dollars):

Digitized by Google

	1919	1920	1921	1922	1923	1924
Jan.....	647	925	958	634	995	1,098
Feb.....	1,075	1,122	946	661	952	1,353
March...	1,135	1,457	1,055	850	1,298	1,632
April.....	1,117	1,544	1,068	1,008	1,322	1,607
May.....	1,158	1,748	1,090	1,129	1,417	1,837
June.....	1,128	1,608	1,084	891	1,381	1,714
July.....	1,349	1,630	1,088	962	1,304	1,716
Aug.....	1,206	1,561	843	1,080	1,369	1,509
Sept.....	1,180	1,478	908	926	1,280	1,410
Oct.....	1,611	1,638	914	1,231	1,648	1,590
Nov.....	1,261	1,630	1,017	1,204	1,614	1,709
Dec.....	1,794	1,461	766	1,071	1,604	1,522
Total..	14,661	17,802	11,737	11,647	16,184	18,697

The per cents which each month's sales are of sales for that year are as follows:

	1919	1920	1921	1922	1923	1924
Jan.....	4.41	5.20	8.16	5.44	6.15	5.87
Feb.....	7.33	6.30	8.06	5.68	5.88	7.24
March....	7.74	8.18	8.99	7.30	8.02	8.73
April.....	7.62	8.67	9.10	8.65	8.17	8.60
May.....	7.90	9.82	9.29	9.69	8.76	9.82
June.....	7.69	9.03	9.23	7.65	8.53	9.17
July.....	9.20	9.16	9.27	8.26	8.06	9.18
Aug.....	8.23	8.77	7.18	9.27	8.46	8.07
Sept.....	8.05	8.30	7.74	7.95	7.91	7.54
Oct.....	10.99	9.20	7.79	10.57	10.18	8.50
Nov.....	8.60	9.16	8.66	10.34	9.97	9.14
Dec.....	12.24	8.21	6.53	9.20	9.91	8.14

It is obvious that no percentages can be selected which will be fully satisfactory. It is probably desirable, therefore, to compute typical percentages by several methods, and finally adopt a compromise between the best of them and the percentages which would result from uniform distribution of sales throughout the year. The table below shows the results of taking the average of the middle two percentages for a given month as representative of that month, of taking average of middle four, or taking average of two latest years. In each case the actual results of the arithmetic are shown, and also results adjusted if necessary to add up to 100, and these adjusted figures smoothed to one decimal place.

Digitized by Google

	Middle Two			Middle Four			Average 1923 and 1924	
	Actual	Adj.	Smoothed	Actual	Adj.	Smoothed	Actual	Smoothed
Jan....	5.65	5.63	5.6	5.66	5.66	5.7	6.01	6.0
Feb....	6.78	6.75	6.7	6.69	6.69	6.7	6.56	6.6
March.	8.10	8.07	8.1	8.17	8.17	8.2	8.38	8.4
April...	8.63	8.59	8.6	8.52	8.52	8.5	8.38	8.4
May...	9.49	9.45	9.5	9.39	9.40	9.4	9.29	9.3
June...	8.78	8.74	8.7	8.60	8.60	8.6	8.85	8.9
July...	9.17	9.13	9.1	8.95	8.95	8.9	8.62	8.6
Aug....	8.35	8.32	8.3	8.38	8.38	8.4	8.26	8.3
Sept...	7.93	7.90	7.9	7.91	7.91	7.9	7.72	7.7
Oct....	9.69	9.65	9.7	9.61	9.62	9.6	9.34	9.3
Nov...	9.15	9.11	9.1	9.23	9.24	9.2	9.55	9.5
Dec....	8.70	8.66	8.7	8.86	8.86	8.9	9.02	9.0

The reason for finding separately the percentages based on the 1923 and 1924 figures is that those two years are both recent and regular. Other years are peculiar in that they contain a number of months in which the percentages are clearly unusual.

It will be noted that the "average of middle four" method gives in nearly every case results intermediate between the results of the other two methods. It therefore seems reasonable to adopt it as a basis. In view of the small number of years on which it is based, however, we cannot feel safe in assuming that future experience will deviate as much from one twelfth of sales in each month as these figures indicate. We may therefore somewhat arbitrarily tone down these percentages by averaging with the series 8.33 in each month, assigning a weight 3 to the "middle four" series, and 1 to the 8.33 series. The following set of percentages results, after smoothing:

Jan.....	6.3	April.....	8.5	July.....	8.8	Oct.....	9.3
Feb.....	7.1	May.....	9.2	Aug.....	8.4	Nov.....	9.0
March.....	8.2	June.....	8.5	Sept.....	8.0	Dec.....	8.7

If a set of indices rather than a set of percentages is desired, it can be obtained by multiplying by 12 before smoothing, and then smoothing the results. The indices are:

Jan.....	76	April.....	102	July.....	106	Oct.....	112
Feb.....	85	May.....	110	Aug.....	100	Nov.....	108
March.....	98	June.....	102	Sept.....	96	Dec.....	105

Digitized by Google

It is impossible to find seasonal factors which will be accurate for single months, but those just found will probably prove fairly satisfactory if current reports must be interpreted by means of seasonal factors derived from only six years' experience.

B 1. The geometric-mean link-relative method. If the data show a trend, it is no longer accurate to use without modification either of the two methods just discussed. One way of meeting this situation is to use one of the link-relative methods — that is to say, to base the seasonal factors on the average ratio of change from one month to the next. Theoretical discussion has not yet shown whether the probable range of error in results obtained by this method is greater or less than with results obtained by other methods. At any rate, the method is convenient and if applied to a reasonably long series — say 10 or 15 years — gives results fairly close to the results of other methods.

Link-relative methods differ in the selection of the average used as typical of the change from month to month. As certain adjustments which have to be made can be justified most readily when the geometric mean is used, we shall regard the geometric-mean link-relative method as fundamental, and the others as modifications.

The procedure is as follows:

Let $a_1^1, a_1^2, a_1^3 \ldots a_1^{12}$ represent sales (etc.) in Jan. Feb. . . . Dec. of the first year, $a_2^1, a_2^2 \ldots a_2^{12}$ in each month of the second year, and so forth. Then the typical ratio of February sales to January sales is

$$\sqrt[n]{\frac{a_1^2\, a_2^2\, a_3^2 \ldots . a_n^2}{a_1^1\, a_2^1\, a_3^1 \ldots . a_n^1}} = {}_1r_2, \text{ say.}$$

Then

$$n \log {}_1r_2 = (\log a_1^2 - \log a_1^1) + (\log a_2^2 - \log a_2^1) + \ldots + (\log a_n^2 - \log a_n^1).$$

Similar equations are true for $\log {}_2r_3$, etc.

To determine the typical ratio of January sales to December sales, we must relate sales in each January to sales in December of the preceding year.

Digitized by Google

We therefore have, denoting by a^1_{n+1} the sales in January in the next year following the n years considered,

$$n \log {}_{12}r_1 = (\log a^1_2 - \log a^{12}_1) + \ldots + (\log a^1_{n+1} - \log a^{12}_n).$$

Now, if the typical ratios chosen are correct, the product of all twelve of them must be equal to *one*, as the ratio of January to itself is *one*.

The sum of their logarithms should therefore be equal to zero. But from the above, we see that

$$n(\log {}_1r_2 + \log {}_2r_3 + \ldots + \log {}_{12}r_1) = \log a^1_{n+1} - \log a^1_1.$$

If we correct each log r by subtracting from it the difference between the logarithms of sales in the first and last Januaries, divided by 12 n, the results, which we shall call ${}_1r'_2$, etc., will have the required property of having their continued product equal to one.

To construct from these r's seasonal indices, take the index for January as k. Then the index for February is clearly ${}_1r'_2$ times this, or ${}_1r'_2k$; the index for March will then be ${}_2r'_3\,{}_1r'_3\,k$, and so forth. But as a matter of convenience, seasonal indices are taken such that their sum is 1200, and therefore if we equate to 1200 the sum of the indices just found in terms of k and the r's, we have an equation by which to determine k. Replacing k by its value thus determined, we obtain a series of indices whose sum is 1200, and which bear to each other the ratios determined as above, with product equal to unity.

The steps in determining these indices are then as follows:

(1) Find logarithms of monthly sales. A four-place table should be used, without attempting interpolation.

(2) Find the successive differences of logarithms, February minus preceding January, March minus February, etc.

(3) Find the arithmetic average of these logarithmic differences for each pair of months. This gives log ${}_1r_2$, etc.

(4) Determine the correction term both by dividing by 12 n the difference of the logarithms of the first and last Januaries,

Digitized by Google

and also by adding the logarithms of the 12 r's and dividing by 12. The agreement of these two is a valuable check.

(5) Subtract this correction from each log r, getting log r'.

(6) Cumulate above results, starting at some convenient point.

(7) Find numbers corresponding to logarithms found in (6).

(8) Add results in (7), and divide by 12.

(9) Divide each result in (7) by the average found in (8).

The work will be illustrated by carrying out the process in detail for the case used as an example of the monthly per cent of annual method.

LOGARITHMS OF $\frac{\text{SALES}}{10{,}000}$

	1919	1920	1921	1922	1923	1924	1925
Jan.....	.8109	.9661	.9814	.8021	.9978	1.0406	1.1501
Feb.....	1.0314	1.0500	.9759	.8202	.9786	1.1313	
March....	1.0550	1.1638	1.0233	.9294	1.1133	1.2127	
April.....	1.0481	1.1886	1.0286	1.0035	1.1212	1.2060	
May.....	1.0637	1.2425	1.0374	1.0527	1.1524	1.2641	
June.....	1.0523	1.2063	1.0350	.9499	1.1402	1.2340	
July.....	1.1303	1.2122	1.0366	.9832	1.1153	1.2345	
Aug.....	1.0813	1.1934	.9258	1.0334	1.1364	1.1787	
Sept.....	1.0719	1.1697	.9581	.9666	1.1072	1.1492	
Oct.....	1.2071	1.2143	.9609	1.0903	1.2170	1.2014	
Nov.....	1.1007	1.2122	1.0073	1.0806	1.2079	1.2327	
Dec.....	1.2538	1.1647	.8842	1.0298	1.2052	1.1824	

The differences of these logarithms, as required in step 2, are:

	1919	1920	1921	1922	1923	1924
Feb.–Jan.....	.2205	.0839	−.0055	.0181	−.0192	.0907
Mar.–Feb.....	.0236	.1138	.0474	.1092	.1347	.0814
April–March..	−.0069	.0248	.0053	.0741	.0079	−.0067
May–April....	.0156	.0539	.0088	.0492	.0302	.0581
June–May....	−.0114	−.0362	−.0024	−.1028	−.0112	−.0301
July–June.....	.0780	.0059	.0016	.0333	−.0249	.0005
Aug.–July....	−.0490	−.0188	−.1108	.0502	.0211	−.0558
Sept.–Aug.....	−.0094	−.0237	.0323	−.0668	−.0292	−.0295
Oct.–Sept.....	.1352	.0446	.0028	.1237	.1098	.0522
Nov.–Oct.....	−.1064	−.0021	.0464	−.0097	−.0091	.0313
Dec.–Nov.....	.1531	−.0475	−.1231	−.0508	−.0027	−.0503
Jan.–Dec.....	−.2877	−.1833	−.0821	−.0320	−.1646	−.0323

Digitized by Google

The details of the remaining steps are shown in the following table:

	Totals	Av.	Av. Minus Corr. Factor	Cumulative	Antilog. × 100	Adjusted (div. by 1.35)
Feb.–Jan.....	.3885	.0648	.0601	.0601	114.8	85 (Feb.)
March–Feb....	.5101	.0850	.0803	.1404	138.2	102 (March)
April–March..	.0985	.0164	.0117	.1521	141.9	105 (April)
May–April....	.2158	.0360	.0313	.1834	152.5	113 (May)
June–May.....	−.1941	−.0324	−.0371	.1463	140.1	104 (June)
July–June.....	.0944	.0157	.0110	.1573	143.6	107 (July)
Aug.–July.....	−.1631	−.0272	−.0319	.1254	133.5	99 (Aug.)
Sept.–Aug.....	−.1263	−.0211	−.0258	.0996	125.8	93 (Sept.)
Oct.–Sept.....	.4683	.0781	.0734	.1730	148.9	110 (Oct.)
Nov.–Oct.....	−.0496	−.0083	−.0130	.1600	144.5	107 (Nov.)
Dec.–Nov.....	−.1213	−.0202	−.0249	.1351	136.5	101 (Dec.)
Jan.–Dec.....	−.7820	−.1303	−.1350	.0001	100.0	74 (Jan.)
Totals.....		.0565			1620.3	
Averages..... (divide by 12)		.0047			135.0	

It may be noted that the maximum difference between these seasonals and the seasonals found by the monthly per cent of annual method is 4 points, in December and in March. On the whole the latter method gives larger seasonals in the later months of the year, because of the fact that trend has not been eliminated, and really is not negligible. The difference, however, is no more than must be expected when seasonals are based on as few as six years.

Some statisticians regard the "correction factor" as a measure of trend. As shown above, it does measure the average monthly per cent increase from the first January to the last, but if these two quantities are out of line with the others, an average rate of change based on them may differ decidedly from the true trend.

B 2. The median or arithmetic-mean link-relative methods. The geometric-mean method is more frequently used in a modified form in which the median of the ratios for consecutive months is selected as typical rather than the geometric mean. The steps in this form are as follows:

Digitized by Google

(1) Find the successive ratios of sales, February divided by January, March divided by February, etc.
(2) Select the median of these ratios for each pair of months.
(3) Find the logarithm of each of these twelve medians.
(4) Cumulate these logarithms, starting at some convenient point.
(5) Determine the correction by dividing by 12 the result of cumulating all 12 logarithms of medians.
(6) Subtract from the cumulated logarithms once, twice, three times . . . up to 12 times the correction term.
(7) Find numbers corresponding to logarithms found in (6).
(8) Add results in (7) and divide by 12.
(9) Divide each result in (7) by the average found in (8).

In step (2), the modified median — i.e., average of middle two, three, four, or more terms is often used, as it gives more stable results than the simple median.

The theoretical justification usually given for this process when the median is used is that the median and geometric mean are often about the same, and therefore the argument outlined above for the geometric mean applies approximately. The reason given for preferring this scheme as a practical procedure is that it is supposed to be shorter. Until it is shown that the time saved is really enough to be significant, the writer recommends the use of the geometric-mean method, because of its direct algebraic justification.

It will be found that if the median L.R. method is applied to the same problem as used in illustrating the G.M. method, the resulting seasonals will differ by as much as three points from the G.M. results and by as much as seven points from the monthly per cent of annual results.

C 1. The method of typical ratio to twelve months' moving average. The fundamental argument back of this method is that an approximate estimate of what sales for the month would be, aside from the effect of seasonal variation, can be obtained by finding the average of the actual sales for that month and for an

Digitized by Google

equal number of months before and after that month. In order to remove completely any seasonal effect, it is desirable to find an average for twelve months.

The steps in computing seasonals by this method are as follows:

(1) Compute the average monthly figure for each successive period of twelve months — i.e., the average for months 1 to 12, then the average for months 2 to 13, then the average for the months 3 to 14, and so on.
(2) Find the average of consecutive pairs of the preceding averages, and allocate each such average to the seventh month of the 13 months on which that pair of averages is based.
(3) Find the ratio of each actual monthly figure to the average allocated to that month.
(4) Select from the ratios for each month a typical value (probably the modified median is best, although the median is often used).
(5) Divide each such typical ratio by the average of the twelve, in order that their sum as adjusted may be equal to 1200.

As the method is straightforward and simple, although somewhat laborious, a detailed working-out of an example will not be given here.

C 2. Method of typical ratio to trend expressed by an equation. A similar but somewhat less time-consuming method is available if a satisfactory trend has been determined in accordance with the methods discussed in the previous chapter. The steps are as follows:

(1) Compute the ordinates of trend for each month. This may require a little fairly obvious algebraic manipulation if the trend is based on annual figures.
(2) Find the ratio of each actual monthly figure to the ordinate of trend for that month.
(3) Divide each such ratio by the ratio of the actual annual figure to the ordinate of trend for the year in which that month occurs.
(4) Select from the resulting quotients a typical value for each

Digitized by Google

of the twelve months (as in C 1, the modified median is probably best, although the arithmetic mean is often to be considered).

(5) Divide each such typical value by the average of the twelve, in order that their sum as adjusted may be equal to 1200.

The great advantage of this method is that it requires very little computation beyond what is done anyway in analyzing a series by means of relation to trend. If the seasonal is to be based on a fairly large number of years, say at least 12, step (3) may ordinarily be omitted without changing the results substantially. If step (3) is included, the correction in (5) will be small, especially if the A.M. is used in step (4).

C 3. Method of typical ratio to a free-hand curve. It has recently been proposed that the ratios on which the seasonals are based should be ratios of actual figures to the ordinates of a free-hand curve drawn by eye in such a way as to represent both trend and fluctuations of the business cycle type. This method, however, depends too much on the judgment of the person drawing the free-hand curve. If several consecutive months generally have low values, the free-hand curve may be brought down each year for those months on the mistaken assumption that the dip is cyclical. If the seasonals were computed on the basis of such a curve, they would not be as small as they should be. In view of this opportunity for personal bias, results computed by this method would never command the confidence which statistical results must command if they are to play their proper rôle in business discussions.

C 4. Method of monthly per cent of annual, modified to eliminate trend. It was noted in discussing method A 2 that that method was not applicable without modification if the series had any appreciable trend. If we divide each ratio of monthly sales to annual, by the corresponding ratio of monthly trend to annual trend, we obtain quotients from which typical values for each month can be selected and adjusted, as in Methods C 1 and C 2. These results will be free of trend and that objection to the "Monthly per cent of annual" Method will be removed. As a matter of fact, this method

Digitized by Google

is then exactly the same in reality as method C 2, as will now be shown:

Let a_{ij} represent sales (or other quantity) in year i and month j;
and A_i represent sales in year i;
and t_{ij} represent the trend for year i and month j;
and T_i represent the trend for year i.

Then the basic element in method C 2, ratio of actual monthly to monthly trend, divided by the ratio actual annual to annual trend, is in symbols:

$$\frac{a_{ij}/t_{ij}}{A_i/T_i}$$

And the basic element in the form of method C 4 just described is

$$\frac{a_{ij}/A_i}{t_{ij}/T_i}$$

It is obvious that these two complex fractions are equivalent, that therefore a determination of seasonal founded on one of them produces the same results as a determination founded on the other.

It is of course a matter of considerable arithmetic to find each ratio t_{ij}/T_i and divide the appropriate percentage by it. If the trend is an exponential curve, the ratio t_{ij}/T_i is, as the reader may readily prove, the same for all values of i, that is to say, for all years. This ratio can therefore be applied at once to the *results* of A 2, to obtain the true seasonal index for month j. In such cases the arithmetic is simple.

In other cases, in which the trend is not an exponential curve, it may be accurate enough to use as a correction factor, to be applied to the A 2 results, the value of t_{ij}/T_i at the middle year. If this process of correction is satisfactory, the arithmetic is much less than in the C 2 method. If the division by t_{ij}/T_i is actually carried through month by month, on the other hand, the C 2 method is preferable because most of the computation would have to be done anyway in other parts of the analysis.

Digitized by Google

Which method should be used in a particular case? No rigid rules can be laid down as to which method will give the best results in a particular case. Sometimes it is desirable to compute the factors by several methods and use those which seem to smooth out best the final ratios to normal with seasonal variation eliminated, or for other reasons seem most satisfactory. If two methods give decidedly different results for certain months, it is desirable to compare the two factors with the original data in detail, noting how many years had actual ratios greater than, and how many less than, each of the two supposedly typical ratios.

While no rigid rule as to choice of method is possible or desirable, the author suggests the following selections as frequently valid:

(*a*) If there is clearly no trend, use method A 2.

(*b*) If there seems to be a trend, but no equation for it has been determined, use method B 1.

(*c*) If an equation for trend has been determined, use method C 2.

Changing seasonal factors. It often happens that conditions change to such an extent that the same seasonal factors do not apply to all years. Several methods have been suggested for meeting this difficulty, but even less definite recommendations can be made than in the case of uniform seasonals. The following lines of attack may be suggested:

(1) Divide the series into several parts and determine separate sets of factors for each part. The division should be based on known changes in conditions or clearly marked differences in the figures.

(2) Instead of taking averages or medians, where required in the methods outlined above, determine a trend.

(3) Instead of taking averages or medians for the entire group of ratios for a given month, find the modified median for the first 7 (or 9 or 5) ratios; then drop the first ratio, add the eighth ratio, and again find the modified median. Find thus a series of typical ratios, each of which is taken as applying

Digitized by Google

to the middle year of the 7 on which it was based. This series of typical ratios may be extended by estimate to cover the last three years and even a year or two into the future. The necessity for estimated extension, however, is the weak point of this method.

Do not hope for extreme accuracy in seasonals. No matter what care is exercised in calculating seasonal factors, it will be found that they will not smooth out all variations apparently due to seasonal conditions. The difficulty is that the effect of seasonal changes is far from uniform, and varies with changes in other conditions. If the seasonal factors are altered to allow for variations which have been noticed for the last two or three years, a fruitless pursuit is undertaken which is known in ballistics as "chasing the error of the gun." This point will be clearer after the later chapters on probability.

One correction which is important in the analysis of current figures allows for the variations in the number of working days in the month. The method of procedure is to express all figures as "average daily sales (production, receipts, etc.)" by dividing each monthly figure by the actual number of working days in that month. The preceding methods, except A 2 and C 4, can then be applied, with obvious adjustment of details.

Summary of Chapter IX.

1. In order to complete the analysis of series of monthly figures, it is necessary to separate out the effect of seasonal variation.
2. There are three purposes for which seasonal factors are obtained:
   (*a*) For use in finding monthly figures with the effect of seasonal variation eliminated;
   (*b*) To show the facts with regard to seasonal variation;
   (*c*) For use in estimating annual on the basis of monthly results.
3. Choice among the seven more or less distinct methods for determining seasonal factors depends on the presence or absence

Digitized by Google

of a trend, desire for simplicity, or accuracy, number of years for which figures are available, and tastes of the computer. It is often desirable to use two or more methods on the same problem and select the factors which work best.

4. For meteorological data and other data in which trend is negligible, the method of monthly means is satisfactory as well as simple. It is particularly adapted to purpose (*b*), showing the facts.

5. If trend is negligible in series of business data, the method of monthly per cent of annual is satisfactory, appropriate, and simple.

6. When the trend is appreciable but has not been expressed in a definite equation, the geometric-mean link-relative method is probably the best available. The median link-relative method is sometimes a satisfactory approximation.

7. The twelve-months moving average is sometimes used as a base of operations in determining seasonal factors. Typical ratios to this base are selected and adjusted. This method is somewhat long, but provides a useful alternative or supplement to the link-relative method in certain cases.

8. The "ratio to trend expressed by an equation" method is very valuable when applicable, and does not require much computation beyond what must be done in analyzing the series by means of a trend and any seasonal factors.

9. The method of "typical ratio to free-hand curve" leaves too much to the judgment of the statistical worker to be valuable in discussions which may be controversial — that is to say, in most statistical work.

10. The method of "monthly per cent of annual," when modified to eliminate trend, turns out to be the same in effect as the "ratio to trend expressed by an equation" method. An approximation is sometimes permissible which requires less computation than the latter method.

11. Changing seasonal conditions can sometimes be allowed for by changing the factors, by methods not yet standardized.

Digitized by Google

12. Extreme accuracy in seasonal factors can never be attained.

Problems. The student can secure large numbers of series by months from the following sources:

(*a*) *The Survey of Current Business*, published by the Department of Commerce.

(*b*) *The Annual Summary*, published by the Standard Statistics Company.

(*c*) Publications of the Harvard Bureau of Business Research.

From these series he can determine seasonal factors by as many different methods as he desires. It is not considered advisable to publish series here, as they would soon become out of date and therefore less interesting. The student, after he has determined seasonal factors for some series which interest him, should apply these factors to current monthly figures as they appear, and compare the conclusions he would draw from such application with the results of the "business man's comparison" of current month with preceding month and same month last year.

Digitized by Google

CHAPTER X

TWO-WAY FREQUENCY DISTRIBUTIONS — TWO-VARIABLE CORRELATION

In the latter part of Chapter VIII, we showed how to determine an equation to summarize the changes in the average value of one variable corresponding to changes in another variable. This scheme is not, however, a complete method of analyzing the relationship between two variables, as it uses averages as a starting point rather than the complete detailed statement of corresponding values. In this chapter we shall go back to the original data, consider various methods of analysis, including that used in Chapter VIII, and then develop further the methods which arise as natural extensions of that method.

Suppose we consider the following table of the ages and educational test scores of 26 girls in the same school grade, with the aim of discovering what relation if any exists between age and performance in the tests:

Girl (by number)	Age	Test Score	Girl	Age	Test Score
1	13.5	91	14	16.5	74
2	14.4	77	15	14.7	72
3	14.5	81	16	14.75	71
4	15.9	48	17	14.75	60
5	15.6	64	18	13.9	68.5
6	14.8	78	19	14.6	66
7	12.7	80	20	15.3	74
8	14.9	67.5	21	14.4	88
9	14.9	80	22	14.8	72
10	14.8	57	23	14.0	82
11	12.6	91	24	14.75	67
12	15.0	59	25	16.7	59
13	14.7	67	26	16.7	48.5

As the figures stand, it is very difficult to see what the general tendency is. As one reads the list, he may note that in a number of cases the younger girls secured high test scores and the older

Digitized by Google

girls low scores. But he cannot be sure that this tendency is generally valid, nor can he put this conclusion into a form which will portray the situation correctly to somebody else, without using some graphic device or summarizing in tabular form.

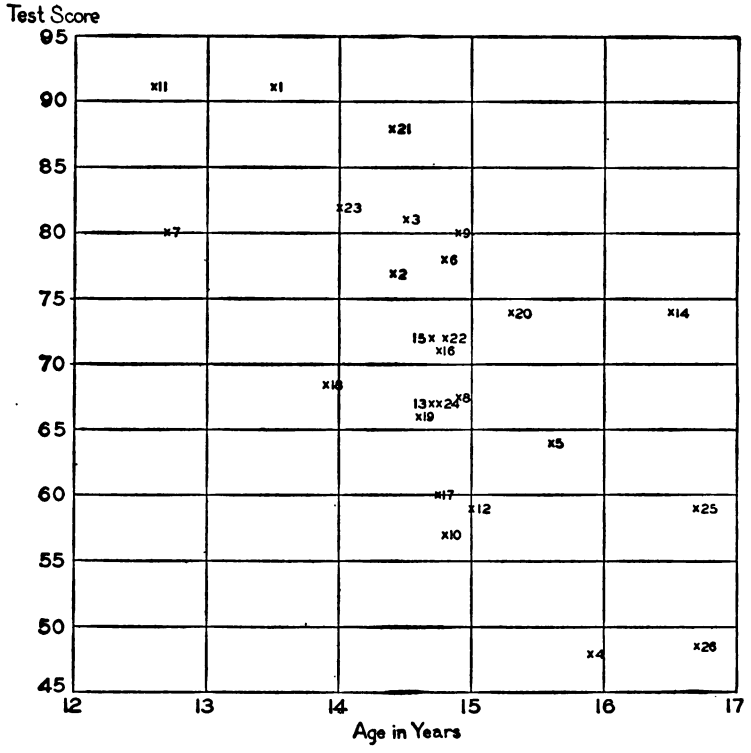


DIAGRAM XXI. SCATTER DIAGRAM OF AGES AND EDUCATIONAL TEST SCORES OF 26 GIRLS IN THE SAME SCHOOL GRADE

A convenient and simple graphic device is the "scatter diagram," constructed as follows: Use the horizontal (x) axis of coördinate paper to show age, and the vertical (y) axis to show test score. Then plot a point for each girl, at the position determined by her age (x) and test score (y). Diagram XXI shows the result when the points for all 26 girls have been plotted. The diagram

Digitized by Google

shows clearly that the general tendency which we suspected really exists, and also gives a visual impression which can be taken in at a glance by any one, both of the existence of this general tendency and of the frequency of exceptions.

For purposes of careful analysis of the relationship, however, and for ready comparison with the relationships which exist between these and other quantities, it is desirable to summarize in a more compact form than the graphic comparison admits. We should have, moreover, some statement in words or figures which can be quoted, as a diagram of this type cannot. As a first step in this direction, then, let us form a table by noting how many girls under 14 had scores over 75, 65 to 75, and under 65, respectively, how many girls of 14 had scores in each of these classes, and how many girls over 14 had scores in each of these classes. The resulting table is as follows:

	Under 14	14	Over 14
Over 75	3	6	0
65 to 75	1	7	2
Under 65	0	2	5

From this table we can see at once that 3 out of 4 girls under 14 scored over 75, while only 6 out of 15 of age 14 and none of the 7 over 14 did so. The percentages of each age group reaching this high standard on the test are therefore 75, 40, and 0, respectively. Similarly, 5 out of 7 girls over 14 failed to reach the score of 65 on the test, while only 2 of age 14, and none under that age, failed to reach that standard.

Another way of summarizing the figures, which might be done roughly from the table, is to find the average score for each of the age groups. The results are:

Age	Average Score
Under 14	83
14	72
Over 14	61

The reader will note that these results constitute a simple functional quantitative series. If the series were a little longer and based on more items, it would be appropriate to determine a trend

Digitized by Google

line for it, as was done at the end of Chapter VIII for the series giving yield of corn in terms of rainfall. In fact the process used in that chapter may be thought of as a late step in the analysis, the earlier steps being the construction of the scatter diagram, and the condensation of the diagram into a table showing the number of years in which each range of yields corresponded to each number of inches of rainfall. Such a table, which is merely a more complete edition of the little table of nine figures above, is known as a two-way frequency table or a correlation table, and is of fundamental importance in the analysis of relationships between two variables.

In order to bring out this importance more clearly, let us examine this table carefully and discuss the various methods which should be used in summarizing it. The table [1] giving the data for the yield-rainfall problem is as follows:

Bushels of Corn per Acre	July Rainfall in Inches								
	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	Total f_y
40-45....		1		1	2	1		1	6
35-40....		1	7	8	7	1			24
30-35....	1	5	8	4	1	1			20
25-30....	1	5	1	2					9
20-25....				1					1
Total f_x	2	12	16	16	10	3	0	1	60

There are a number of things about this table we should note before we begin to draw numerical consequences. In the first place, we must always remember that it is a tabular presentation of a scatter diagram, formed by dividing the entire field into equal rectangles and counting the number of dots in each rectangle. While for simplicity and ease of computation, we may think of all the points as having the values of the variables at the mid-point of the rectangle, it would always be more accurate to go back to the actual values of x and y for the individual points.

In the second place, a brief consideration of the table shows us

[1] Slightly modified from table on page 421 of A. S. Gale and C. W. Watkeys' *Elementary Functions*. Henry Holt & Co., New York, 1920.

Digitized by Google

that the relation between yield and rainfall is not uniform and exact, but irregular, and valid only as an expression of general tendency. Such relations are characteristic of the natural and social sciences as distinguished from the physical sciences. Where we are not able to control other factors than the two which we want to compare, a direct relationship which actually exists may be so obscured that it appears only in the average of a considerable number of cases. We see above, for instance, that even when the rainfall was only two to three inches, the yield was sometimes large. We can explain this by supposing that in such years other factors than July rainfall were unusually favorable. In the physical sciences, we could arrange conditions so that these other factors would be held constant while we observed the correspondence between the two in which we are interested. If we desire to measure the relation between the weight hung at the middle of a beam supported at its two ends and the deflection of the beam, we arrange to keep the beam at a constant temperature, to keep away drafts, and to keep the length of the beam and the point of application of the weight rigorously unchanged; we then put on a succession of weights and measure the resulting deflection very carefully several times for each weight. Conditions would be more nearly analogous to the above agricultural problem if the experiment were made in a room with a gale blowing through, with the distance between supports changing by unknown amounts between experiments, and the weights applied at points sometimes to one side and sometimes to the other of the middle of the beam, the exact amount of displacement being unknown.

Readers trained in the exact sciences will be inclined to throw up their hands in holy horror at this point. If the alternative actually available to the statistical process were carefully controlled experiment, the writer would agree with this point of view. But the alternative actually available is guessing or a subjective process of analysis known as "judgment." The statistical process is therefore the most scientific that is available when controlled experiment is impossible. Care should be exercised in connection with this process to eliminate the effect of extraneous factors as far as possible, and to require a considerable number of items as a basis for

Digitized by Google

drawing any conclusions, in order that peculiar items may be counterbalanced by items of opposite characteristics or may be rendered innocuous by inclusion of other more nearly normal items.

As one method of summarizing the above yield-rainfall table, let us condense it into a table with nine "cells," like the one used to summarize scatter diagram at the beginning of this chapter. The result is as follows:

	Less than 3 Inches	3-5 Inches	Over 5 Inches
Over 40	1	1	4
30–40	7	27	10
Under 30	6	4	0

The results shown in the table may be stated by the use of percentages as follows:

When the July rainfall was less than three inches, the yield was less than 30 bushels per acre in 43 per cent of the years; when the rainfall was between three and five inches, the yield was less than 30 bushels per acre in 12.5 per cent of the cases, and when the rainfall was over five inches the yield was never less than 30 bushels per acre. We may also note that the rainfall exceeded five inches in only 23 per cent of the years, but that it exceeded five inches in two thirds of the years in which the yield was over 40 bushels per acre.

We have noted, then, four methods of finding or showing the relation between two characteristics:

(*a*) Constructing a scatter diagram;

(*b*) Consolidating the facts into a two-way frequency table;

(*c*) Consolidating still further into a table with nine cells;

(*d*) Stating the key percentage relations.

Yet another way of showing the facts is to construct a model of the two-way frequency table in three dimensions, by erecting a column above each cell of the table and perpendicular to its plane, the height of the column being proportional to the frequency. The difficulty of model construction in three dimensions makes this method rather unusual in actual statistical work, but the idea of such a model is an important aid in understanding two-way tables. One way of avoiding the practical difficulty of construction and re-

Digitized by Google

production of three-dimensional models is to draw the contour lines for the surface. Just as map-makers draw lines on the map connecting points of equal elevation, so the statistician may draw, on the grid for the two variables, lines connecting points for which the frequency is the same. While the exact location of such lines in a particular case calls for trained statistical judgment, reenforced with technical skill, it is a simple matter, but important, to understand their significance when drawn.

The method of contour lines has been used by Sir George H. Knibbs, the Australian statistician, in his *Report on the Mathematical Theory of Population.* In one diagram, for instance, he gives the contour lines of what he calls the "Protogamic Surface" — that is to say, the surface showing the two-way frequency distribution of ages of husbands and wives at the time of marriage.

In the opinion of the author, this method will be used with increasing frequency as more refined data become available and more statisticians become familiar with the required technique.

The method of averages. Another method of summarizing the two-way table, which seems natural to almost everyone who has to deal with problems in which such tables arise, is to find the average yield for each rainfall, and the average rainfall for each classification of yield. We easily find from the table these averages, as follows:

Rainfall	Average Yield	Yield	Average Rainfall
1–2 in.	30. bu.	20–25 bu.	4.5 in.
2–3	31.7	25–40	2.94
3–4	34.4	30–35	3.6
4–5	34.4	35–40	4.5
5–6	38.0	40–45	5.5
6–7	37.5		
8–9	42.5		

The first series, we recognize at once, is the functional quantitative series whose equation was found in Chapter VIII to be:

$$y = 35 + 1.7\,(x - 4.5),$$ where y = yield and x = rainfall.

The second series is a functional quantitative series of the same

Digitized by Google

sort as the first, and by the same method as used for the first series its equation may be found to be:

$$x = 3.8 + .15(y - 32.5).$$

These two equations, known as the *regression equation of y on x* and of *x on y*, respectively, summarize the relationship between x and y from two different points of view. The former gives the best means of estimating yield for a year in which we know the rainfall; the second gives the best equation for estimating rainfall when we know the yield. These equations, however, do not in this form summarize the degree of the relationship between x and y in a single figure.

The coefficient of correlation. A single figure summarizing the degree of relationship, however, has been derived from the formulas for the two regression equations. The formulas are:

$$\text{(I)} \quad y - \bar{y} = \frac{\Sigma f\,(x - \bar{x})\,(y - \bar{y})}{\Sigma f\,(x - \bar{x})^2}\,(x - \bar{x})$$

$$\text{(II)} \quad x - \bar{x} = \frac{\Sigma f\,(y - \bar{y})\,(x - \bar{x})}{\Sigma f\,(y - \bar{y})^2}\,(y - \bar{y})$$

Formula II has not been stated above, but is merely the result of interchanging x and y in the formula for the regression equation as proved originally (formula VIII–2). It will be noted that the numerator of the second regression coefficient (x on y) is the same as the numerator of the first regression coefficient (y on x).

The denominators of these coefficients may be stated as $N\sigma_x^2$ and $N\sigma_y^2$, respectively, introducing the subscripts x and y in the formula VII–2. Now it was suggested in Chapter VII that in certain problems, it is convenient to use *standardized* deviations; that is to say, to divide the actual deviations by the standard deviations. On the right-hand side of equation I, we have σ_x^2 in the denominator of the coefficient, and can transfer one of the σ_x's to divide the $x - \bar{x}$, to get $(x - \bar{x})/\sigma_x$, that is to say, the standardized deviations of x. Then dividing both sides of the equation by σ_y, it takes the form:

Digitized by Google

$$\frac{y-\bar{y}}{\sigma_y} = \frac{\Sigma f(x-\bar{x})(y-\bar{y})}{N\sigma_x\sigma_y} \cdot \frac{x-\bar{x}}{\sigma_x}$$

The coefficient of $(x-\bar{x})/\sigma_x$ in the equation in this form is symmetrical as respects x and y and is not affected by the units used in measuring x or y, as both x and y occur to the same power in numerator and denominator. Exactly the same coefficient will result if we manipulate equation II above as we did equation I, stating it in the form of standardized deviation of x equals a coefficient times a standardized deviation of y. This form of coefficient is therefore more general and can be understood apart from the context better than the unmodified regression coefficients. It is therefore used as a general measure of degree of relationship and is known as the *coefficient of correlation.* The formula is:

$$\text{Coefficient of correlation } (r) = \frac{\Sigma f(x-\bar{x})(y-\bar{y})}{N\sigma_x\sigma_y}$$

or

$$\frac{\Sigma f(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma f(x-\bar{x})^2 \times \Sigma f(y-\bar{y})^2}}$$

The student should note that our derivation of this formula is based on the formula for a linear trend line. If the given data are not well summarized by such a trend line, then the coefficient of correlation is in so far inadequate as a summary of degree of relationship. In fact, it may in certain cases be very misleading.[1]

Proof that r always lies between $+1$ and -1. By carrying through a step or two further some of the algebra used in deriving the coefficient of regression, and incorporating in the work the symbols r as defined above, σ_x and σ_y, we shall now show that r must lie between -1 and $+1$, in every case of practical importance for the statistician.

In Chapter VIII, the regression coefficient was derived by minimizing the expression $\Sigma(Y-mX)^2$, where Y and X stand for $y-\bar{y}$ and $x-\bar{x}$ respectively. It was found that

[1] Cf. article by H. L. Rietz, "On Functional Relations for which the Coefficient of Correlation is Zero," in the *Quarterly Publications of the American Statistical Association*, Sept., 1919, no. 127.

Digitized by Google

$$m = \Sigma f(x - \bar{x})\,(y - \bar{y})/\Sigma f(x - \bar{x})^2$$

Using the above definition of r, we have

$$m = Nr\sigma_x\sigma_y/N\sigma_x^2 = r\sigma_y/\sigma_x$$

The expression which is made a minimum by this value of m is then

$$\Sigma\left(Y - r\frac{\sigma_y}{\sigma_x}X\right)^2$$

or, expanding,

$$\Sigma Y^2 - \frac{2r\sigma_y\Sigma XY}{\sigma_x} + \frac{r^2\sigma_y^2\Sigma X^2}{\sigma_x^2}$$

Simplifying, we have

$$N\sigma_y^2 - \frac{2r\sigma_y Nr\sigma_x\sigma_y}{\sigma_x} + r^2\sigma_y^2 \times N = N\sigma_y^2 - Nr^2\sigma_y^2 = N\sigma_y^2(1 - r^2)$$

But as this expression is the sum of squares, which are not all zero in the practical case, it must be positive; therefore r cannot be numerically greater than 1, that is to say, must lie between $+1$ and -1. For r to be exactly $+1$ or -1, $1 - r^2$ must be zero and therefore each term of the type $Y - r\frac{\sigma_y}{\sigma_x}X$ exactly equal to zero. This can happen only when every observed Y is the same multiple of X, which is a type of relationship whose study is not facilitated by statistical methods. It is therefore statistically adequate to say that r must always lie between -1 and $+1$, with no mention of the possible equality.

Second method of derivation of formula for coefficient of correlation. As it is desirable to approach the concept of correlation from several different angles, the following alternative method for deriving the formula may be of interest.

Let us try to measure directly the degree of relationship between two sets of quantities, x and y. A little experimenting will show that any measure constructed from the quantities as they stand is very difficult to interpret. A change of one unit in x, for instance, may be a large proportion of the greatest change that ever occurs

Digitized by Google

in that quantity, while a change of one unit in y may be insignificant. In order to reduce x and y to comparable scales of measurement, then, it might well seem expedient to divide them by their respective standard deviations. A further step towards putting them on a comparable basis would be to use the deviations from the means rather than the quantities themselves. That is to say, we might well use in our search for a measure of correspondence the standardized deviations $(x - \bar{x})/\sigma_x = X'$, say, and

$$(y - \bar{y})/\sigma_y = Y'$$

If we had perfect correspondence, we might hope to have every difference $X' - Y'$ equal to zero. To get a measure of failure to correspond, we might take the average of the squares of these, that is to say,

$$\Sigma(X' - Y')^2/N$$

As clearly $\Sigma X'^2 = \Sigma Y'^2 = 1$, this "average failure to correspond" can be simplified to

$$1 - 2\frac{\Sigma X'Y}{N} + 1 = 2 - 2\frac{\Sigma X'Y'}{N}$$

It is clear that the significant term here is $\Sigma X'Y'/N$, which may be called the measure of degree of correspondence. Substituting for X' and Y' their values, this measure is

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{N\sigma_x\sigma_y}.$$

or exactly r as found above.

Significance of various values of r. A clear idea of the interpretation to be placed on r in a given problem depends on knowing how large r has been in other problems statistically similar to the given problem. In general r is a number which measures (sometimes somewhat imperfectly) the extent to which large values of x are associated with large values of y, and small values of x with small values of y. If such association were perfect in any given case, the deviation of each y from the mean y would be a constant positive

Digitized by Google

multiple of the deviation of x from the mean x, and r would be equal to $+1$. If large values of x were invariably associated with small values of y, and small values of x with large values of y, the deviation of each y from the mean y would be a constant negative multiple of the deviation of x from the mean x, and r would be equal to -1. If there is no connection at all between x and y, r is equal to zero. An intermediate value of r indicates a relationship between these extremes.

As an aid in interpretation, we may note that historical series, correlated without first eliminating trends, sometimes yield values of r over .95. Ages of husband and wife show a correlation of over .9, if taken over a large group with a wide range of ages. On the other hand, very much smaller values of r are found to indicate a significant degree of correlation in psychological or educational statistics, where the problem is, for instance, to find the degree of correspondence between scores on a psychological test and success in college work. A standard statement of the interpretation of such results is that of Rugg,[1] who regards correlation as "negligible" or "indifferent" when r is less than .15 or .20; as being "present but low" when r ranges from .15 or .20 to .35 or .40; as being "markedly present" or "marked" when r ranges from .35 or .40 to .50 or .60; and as being "high" when it is above .60 or .70.

It will be found that r will tend to run higher in a group whose members differ decidedly in the two characteristics which are being correlated, than in a group whose members are more nearly homogeneous in those respects. The correlation between intelligence test scores, and arithmetical ability, for instance, would be expected to be higher in a group taken from the community as a whole than in a college Freshman class. While this is a little difficult to prove rigorously, it may be understood by inspection of the formula for r. The key to this formula is the value of the average product-sum, which, as an arithmetic mean, is very much affected by extreme cases. And extreme cases of very low intelligence combined with very poor arithmetical ability are likely to be included in the general group but excluded from the college group.

Because of the vagueness in the interpretation of r brought out by

[1] *Statistical Methods Applied to Education*, by H. O. Rugg, p. 256.

Digitized by Google

the preceding discussion, r should not be relied on to give the complete and only analysis and summary of relationships in a given problem. Other schemes, such as those discussed earlier in the chapter, are valuable supplements, and the use of the two regression equations, with attention to their degree of linearity, is to be recommended.

Computation of coefficients of correlation and regression. While we have defined and discussed the coefficients of regression and of correlation largely from the point of view of the functional quantitative series, that point of view does not necessarily indicate what is the easiest method of computing these coefficients. It is evident from the formulas that the only computation not fully discussed in previous chapters is that of the average product sum, i.e.,

$$\Sigma fXY/N = \Sigma f(x - \bar{x})\,(y - \bar{y})/N$$

which for brevity we shall call p. Then in this notation, m_x, the coefficient of regression of y on x, and m_y, the coefficient of regression of x on y, and the coefficient of correlation r, are given by the formulas:

$$m_x = p/\sigma_x^2,\ m_y = p/\sigma_y^2, \text{ and } r = p/\sigma_x\sigma_y \quad \text{(Formulas X–5)}$$

If, then, we can devise any convenient method of computing p, the computation of the three coefficients is a simple matter. The method of Chapter VIII for computing ΣXY (i.e., Np) is satisfactory when the data are in the form of a quantitative series, but can be simplified when the data are simply ungrouped pairs of values of x and y, and can with advantage be replaced with another method when the data are in the form of a two-way frequency table. These variant methods will be taken up shortly.

Formula VIII–3 may conveniently be expressed in the above "p" notation, by letting $\Sigma fX'Y'/N$, the average product-sum referred to assumed means, be denoted by p'. We then have:

$$\Sigma f\,(x - \bar{x})\,(y - \bar{y}) = \Sigma fX'Y' - Ncd \quad \text{(Original Formula VIII–3)}$$

Dividing by N,

$$\Sigma fXY/N = \Sigma fX'Y'/N - cd \text{ or } p = p' - cd \quad \text{(Formula X–6)}$$

Digitized by Google

An obvious corollary of formulas X–5, to which attention should be called, is that

$$\sqrt{m_x m_y} = p/\sqrt{\sigma_x^2 \sigma_y^2} = p/\sigma_x \sigma_y = r$$

Computation of r for a simple case of pairs of values of x and y. Merely for illustration of the formulas, suppose we desire the correlation between two quantities when the pairs of corresponding values are (3, −2) (5, 1) (6, 1), (7, 2) and (9, 3). Calling the first quantity x and the second y, the computation is as follows:

	x	y	x^2	y^2	xy
	3	−2	9	4	−6
	5	1	25	1	5
	6	1	36	1	6
	7	2	49	4	14
	9	3	81	9	27
Totals...	30	5	200	19	46
Av.....	6	1	40	3.8	9.2

$\sigma_x^2 = 40 - 36 = 4$

$\sigma_y^2 = 3.8 - 1 = 2.8$

$p = p' - cd$

$= 9.2 - 6 = 3.2$

$\sigma_x = 2$

$\sigma_y = \sqrt{2.8} = 1.673$

$$r = \frac{3.2}{2 \times 1.673} = \frac{3.2}{3.346} = .956$$

Computation of coefficients of regression and correlation from two-way frequency table. As the computation of p is most easily done by a considerably different method if the data are given in the form of a two-way frequency table, it is desirable to show this method in detail. We shall use for this purpose a table consolidated from one given in the March, 1923, issue of the *Journal of the American Statistical Association*, in an article by Harry A. Wembridge entitled "Experiment and Statistics in the Selection of Employees." Various methods were tried at a certain factory to test candidates for employment before hiring them. The results of various methods were compared by finding the coefficients of correlation between the ratings on the tests and a definite measure of subsequent factory efficiency which was available — the per cent which an employee's performance was of the standard set for that job. It turned out that the method of testing which gave the highest correlation with factory success was a mental test. The following table summarizes the results:

Digitized by Google

NUMBER OF INDIVIDUALS SECURING SCORE ON MENTAL TEST INDICATED AT LEFT AND REACHING PER CENT OF STANDARD PERFORMANCE INDICATED AT HEAD OF COLUMNS

POINTS ON MENTAL TEST (y)	PER CENT OF STANDARD REACHED (x)								
	50–65	65–80	80–95	95–110	110–125	125–140	140–155	155–170	Total
60–69...						2	2		4
50–59...		6	15	18	8	6	5	5	63
40–49...	8	32	30	29	9	2	4	2	116
30–39...	13	18	33	13	4	1			82
20–29...	6	9	5	1					21
10–19...	3	1							4
Totals...	30	66	83	61	21	11	11	7	290

On page 212 this table is shown with all the computation needed to determine the means, the standard deviations, the regression equations, and the coefficients of correlation. The steps are as follows:

1. As was done in computing the means and the deviations by the deviation method, select for x and for y a class interval whose mid-point may conveniently be used as the assumed mean. We usually try to select this interval so that the true mean will fall in it, but sometimes the computations involve smaller numbers if the interval chosen is more nearly in the middle of the range. In this case, the choice of the interval 40–49 is clearly best for y, the interval 80–95, as used for x, is satisfactory; in the latter case, however, 95–110 would have been a little more central, and on the whole about as good.

2. The computation of the true mean and of the standard deviation of y is carried out exactly as in previous chapters. The same thing is done for x, except that the work is done below the table, in rows, instead of to the right in columns. The means and deviation are computed from the totals to the right and at the bottom of the original table, without any use of the details in the body of the table.

Digitized by Google

Computation of Coefficients of Correlation and Regression for Twofold Frequency Table

NUMBER OF INDIVIDUALS SECURING SCORE ON MENTAL TEST INDICATED AT LEFT AND REACHING PER CENT OF STANDARD INDICATED AT HEAD OF COLUMNS

Points on Mental Test (y)	Per Cent of Standard Reached (x)									$\frac{Y'}{\frac{y-44.5}{10}}$	$f_y \times Y'$	$f_y \times Y'^2$	$f_y X'Y'$
	50–65	65–80	80–95	95–110	110–125	125–140	140–155	155–170	f_y				
60–69						2^{6}_{12}	2^{3}_{16}		4	2	8	16	28
50–59		6^{-1}_{-6}	15	18^{1}_{18}	8^{3}_{16}	6^{2}_{18}	5^{4}_{20}	5^{5}_{25}	63	1	63	63	91
40–49	8	32	30	29	9	2	4	2	116	0	0	0	0
30–39	13^{2}_{26}	18^{1}_{18}	33	13^{-1}_{-13}	4^{-2}_{-8}	1^{-3}_{-3}			82	−1	−82	82	20
20–29	6^{4}_{24}	9^{3}_{18}	5	1^{-2}_{-2}					21	−2	−42	84	40
10–19	3^{3}_{18}	1^{2}_{8}							4	−3	−12	36	21
f_x	30	66	83	61	21	11	11	7	290		−65	281	200
$X' = \frac{x-87.5}{15}$	−2	−1	0	1	2	3	4	5		(Av.)	−.224 (d)	.9690 (σ'^2_y)	.6897 (p')
$f_x \times X'$	−60	−66	0	61	42	33	44	35	89	.307 (c)			
$f_x \times X'^2$	120	66	0	61	84	99	176	175	781	2.6931 (σ'^2_x)			
$f_x X'Y'$	68	33	0	3	8	27	36	25	200				

$c = .307$ class-intervals $= .307 \times 15$ per cent $= 4.60$ per cent; $\bar{x} = 87.5 + 4.60 = 92.10$.

$d = -.224$ class-intervals $= -.224 \times 10$ points $= -2.24$ points; $\bar{y} = 44.5 - 2.24 = 42.26$.

$\sigma_x^2 = \sigma'^2_x - c^2 = 2.6931 - (.307)^2 = 2.6931 - .0942 = 2.599$; $\sigma_x = 1.612$ class-intervals $= 24.18$ per cent.

$\sigma_y^2 = \sigma'^2_y - d^2 = .9690 - (.224)^2 = .9690 - .0502 = .9188$; $\sigma_y = .9585$ class-intervals $= 9.585$ points.

$p = p' - cd = .6897 + .224 \times .307 = .6897 + .0688 = .7585$; $r = p/\sigma_x\sigma_y = .7585/(.9585 \times 1.612) = .491$.

$m_x = p/\sigma_x^2 = .7585 \times 10 \times 15/2.60 \times 15 \times 15 = .1945$; $m_y = p/\sigma_y^2 = .7585 \times 10 \times 15/.9188 \times 10 \times 10 = 1.238$. $r = \sqrt{m_x m_y} = .491$.

Digitized by Google

3. It is convenient to carry out most of this work in terms of "class-intervals" and multiply by width of interval at the last stage; it is important to remember that class-intervals are being used, if this short-cut is employed. Some workers prefer to avoid confusion by sticking to the original units at all times.

4. The computation of the average product-sum p may be done in three ways, two of which are shown in the table. In each of these methods, we write in each cell of the table the value of the product $X'Y'$ (in class-intervals) for that cell. This product may well be written in small figures above and to the right of the figure for frequency in that cell. For two of the methods, we write also in the cell, to the right of the figure for frequency but below it, the product of frequency into $X'Y'$. The first method of getting p' is then to add these products by rows (giving last column to right) and the second to add them by columns (bottom row of table, $f_x X'Y'$). The agreement of these two totals checks the addition.

The third method for getting p' checks the multiplication of frequency into product $X'Y'$. We take the sum of all the frequencies for cells in which the product $X'Y'$ is zero; then the algebraic sum of all the frequencies for which the product $X'Y'$ is 1 or -1, transferring the minus sign if present from the product to the frequency. We write down both the sum of the positive and the sum of the negative frequencies, in order to be able to check that we have included all items. Similarly we add up the frequencies for cells in which the product $X'Y'$ is 2, 3, 4, and so on. We then multiply each $X'Y'$ by the corresponding net frequency and total the products, thus deriving Np' by a different method. The work for this problem is shown below.

$X'Y'$	Frequency (F)	$FX'Y'$
0	116 + 53 = 169	0
1	36 − 19 = 17	17
2	30 − 5 = 25	50
3	7 − 1 = 6	18
4	11 = 11	44
5	5 = 5	25
6	5 = 5	30
8	2 = 2	16
Total	265–25	200

Digitized by Google

The totals show that we have utilized 265 positive and 25 negative frequencies, thus accounting for the entire 290 items, and that $Np' = 200$, as found above.

5. It is convenient to divide $\Sigma fY'$, $\Sigma fY'^2$, $\Sigma fX'Y'$, $\Sigma fX'$, and $\Sigma fX'^2$ by N (Σf_x or Σf_y) and place the result immediately below or to the right of the summation. These results are the c, $\sigma_y'^2$, p', c, and $\sigma_x'^2$ required for further computations.

6. The computations at the bottom of the page show the derivation of the true means, standard deviations, and coefficients of correlation and regression from the results mentioned in Step 5 and the assumed means. It should be noted that the results found above in terms of class-intervals must be converted into "points" or "per cents" before they are substituted in the formulas for m_x and m_y. This conversion is accomplished by multiplying p by the product of the widths of the x and y intervals, σ_x^2 by the square of the width of the x interval, and σ_y^2 by the square of the width of the y interval.

7. The regression equations are:

$$y - 42.26 = .1945\,(x - 92.1) \text{ or } y = 43.8 + .1945\,(x - 100)$$

and $$x - 92.1 = 1.238\,(y - 42.26) \text{ or } x = 89.3 + 1.238\,(y - 40)$$

The second form of each equation will prove more convenient than the first, for determining the average y corresponding to a given x, or the average x corresponding to a given y.

For most purposes, these figures should be vigorously rounded off before being used, and the equations stated as:

$$y = 44 + .2\,(x - 100) \text{ and } x = 89 + \frac{5}{4}\,(y - 40)$$

In words, the latter equation states that the typical individual who scores 40 on the mental test will reach 89 per cent of standard in factory performance, and that each improvement of four points in the mental score indicates an improvement of five per cent in performance. Similarly, each drop of four points in mental score indicates a drop of five per cent in performance.

Digitized by Google

8. Not only the regression coefficients, but also the other constants, should be rounded off after the computations are completed. Criteria will be given later, after the discussion of probability and the theory of sampling, as to the degree of reliability of the results of such computations. At present we suggest that if the computation is based on less than one hundred items, little importance should be attached to the third or later significant digits in the constants; if the computation is based on more than one hundred but less than five hundred items, the third significant digit [1] is of value if the first digit is 1 or 2, but not otherwise.

9. The first regression equation might also be found by finding the average y for the items in each column and fitting an equation to the functional quantitative series so formed. It will be found that average y's must be obtained accurate to at least two decimal places, if the results of such a computation are to agree with the results just found.

10. The method of computation explained above is not quite as accurate as the method of determining x^2, y^2, and xy individually for each point and adding, as explained in a previous section. The difficulty is that the use of the two-way table is based on the assumption that all the items in each cell may be treated as if concentrated at the mid-point of the cell. The resulting error is known as "the error of grouping." This error is diminished by increasing the number of cells. In certain cases this error may be corrected, so far as the standard deviations are concerned, by subtracting $w^2/12$ from σ^2, where w is the width of the class-interval. In any case this correction formula can be used to show about how large an inaccuracy may result from any contemplated grouping; if the inaccuracy so indicated is larger than is permissible for the purpose in hand, a larger number of cells is desirable. Any increase in the number of cells of course increases the amount of arithmetic work, and the method of determining x^2, y^2, and xy individually is prohibitively laborious if there are more than about 60 items.

11. The regression coefficients may be checked roughly as fol-

[1] The "third significant digit" in a number is the third from the left, if we begin to count as soon as we strike a digit not zero. For instance, in 1.238, 3 is the third significant digit; in 0.003872, 7 is the third significant digit.

Digitized by Google

lows: by inspection of the first column, we see that a score of 33 corresponds to about 60 per cent of standard; from the next to the last column (the last column does not seem to carry through the trend of the table) a score of 53 corresponds to about 150 per cent of standard. An increase of 90 per cent of standard leads to an increase of about 20 in the score, or an increase at the rate of about .22 points for each increase of one per cent in performance.

Similarly, a per cent of standard of about 68 corresponds to a score of about 25 points, and 109 per cent of standard corresponds to a score of about 55. An increase of about 30, then, in score leads to an increase of about 41 per cent in performance, or an increase of about 1.4 per cent for each increase of one point in test score.

These coefficients are both high as compared with the .2 and 1.24 found by actual computation, but nevertheless are a good enough check to catch errors of over 25 per cent. If the scatter diagram is plotted, a graphic determination of slope by inspection may serve the same purpose.

Use of *r* for determining lag of one historical series as compared with another. An important application of the coefficient of correlation to economic statistics is to determine the length of time by which the fluctuations of one series precede those of another series. This method will be explained by means of an example taken from the July, 1923, *Business Bulletin* of the Cleveland Trust Company. The problem considered was, to find the relation between variations in industrial stock prices and those of general business. The variables actually used, as shown in the accompanying graph (Diagram XXII), were the variations from estimated normal of general business and the prices of industrial stock from 1900 through the first half of 1923.

A glance at the graphs of these quantities is enough to show that the stock line has almost invariably moved earlier than the general business line. At times the general business line follows the stock line within two months; at other times the interval seems to be almost a year. Sometimes the correspondence between the two series is close, but at other times one shows ups and downs not

Digitized by Google

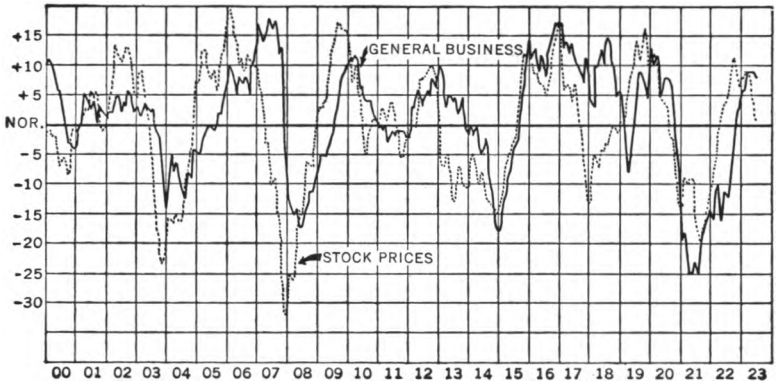


DIAGRAM XXII. VARIATIONS FROM NORMAL OF GENERAL BUSINESS AND THE PRICES OF INDUSTRIAL STOCKS FROM 1900 TO 1922

duplicated in the other. We get a measure of the typical interval between changes in the two series, and of the degree of correspondence, by using the coefficient of correlation as follows: Compute the coefficient of correlation pairing together the business and stock indices for each month. The result turns out to be .58, high enough to indicate significant correlation, but not really close correlation.

Let us next compute r by pairing each stock figure with the general business figure which actually came two months later. That is to say, speaking graphically, we move the general business line back two months, or the stock line forward two months. The coefficient of correlation then turns out to be .67, considerably higher. Moving the stock line ahead four months, we find $r = .71$; five months, $r = .84$; six months, $r = .72$. We therefore conclude that the closest connection is that between the stock figures and the general business figure of five months later, and that the connection is moderately close.

This general relationship can therefore be used for predicting general business conditions on the basis of stock prices, and is in fact so used in the Bulletin quoted. Considerable use has already been made of this method of determining the lag of one series as compared with another, and it seems to promise valuable additional results in the study of the business cycle and of the rela-

Digitized by Google

tions between the various general series and the sales or other series of a particular business.

In using results derived by this method, we must always bear in mind that they merely summarize what has most frequently or on the average happened in the past. They are the best basis we have for judging what is most likely to happen in the future, but must be modified or used with caution when underlying conditions have changed. They are not based on permanent physical or astronomical causes such as the forces of gravity and elasticity which produce the vibrations of a pendulum or of a piano string, and the ocean tides. They are based rather on the interactions of an unanalyzed mass of psychological, physiographic, legal, political, and other factors, many of which are subject to irregular changes.

PROBLEMS ON CHAPTER X

1. Number of men six feet tall who fell in each age and weight group.

Weight	Age					
	10–19	20–29	30–39	40–49	50–59	60–69
103–117	1	1				
118–132	1	10	7	4	4	1
133–147	47	623	289	105	39	9
148–162	90	1879	1294	368	122	14
163–177	53	1804	1628	680	197	22
178–192	11	799	1131	532	192	28
193–207	2	269	659	406	126	15
208–222	1	122	334	235	93	11
223–237		32	126	91	33	4
238–252		8	51	29	16	2
253–267		3	4	4	1	
268–282			2		1	

Find the coefficient of correlation and both regression equations.

2. If the distribution of weight for the women of example 1, Chapter VIII, was as follows, determine the value of r.

Weight	88–112	113–137	138–162	163–187	188–212	213–237
No. of women	2	47	46	16	3	1

3. The following table shows, for a group of 212 teachers in cities of a certain class, the number in each age and salary group. Find both regression equations and the coefficient of correlation. (Figures adapted from Report of National Commission on the Emergency in Education.)

Digitized by Google

Salary	Age									Total
	18–23	24–29	30–35	36–41	42–47	48–53	54–59	60–65	66–71	
$450– 549.....				1						1
550– 649.....	4	2								6
650– 749.....	15	17	3	1						36
750– 849.....	9	21	4	9	4	1				48
850– 949.....		15	3	5	1	2	1			27
950–1049....		12	20	10	6	3	3	3		57
1050–1149....		1	7	13	8	1		2	1	33
1250–1349....			1			1	1			3
1450–1549....			1							1
Total.....	28	68	39	39	19	8	5	5	1	212

The correlation ratio. Another measure of correlation, which is not dependent for its value on the linearity of the regressions, as the coefficient of correlation is, is the *correlation ratio.* This is based on the comparison between the means of the columns (or rows) with the general mean. It is clear that if there is no correlation, the column-means should except for accidental errors be the same as the general mean. The differences between these column-means and the general mean are therefore measures of connection between x and y, and some sort of combination of these differences may be taken as a measure of correlation. As these differences are a kind of deviation from the mean, it is natural to combine as we do in finding the standard deviation — i.e., by squaring, multiplying by frequency, adding, dividing by N, and extracting square root. Further, as the result is a sort of deviation in the y's, it is reasonable to standardize by dividing by σ_y. The formula then is, denoting this new measure by η, and the mean of the k'th column by $\bar{y}_k$

$$\text{Correlation ratio} = \eta_y = \frac{\sqrt{\Sigma f\,(\bar{y}_k - \bar{y})^2/N}}{\sigma_y} = \sqrt{\frac{\Sigma f\,(\bar{y}_k - \bar{y})^2}{\Sigma f\,(y - \bar{y})^2}}$$

There is similarly another correlation ratio η_x, based on the means of the rows as compared with the general mean of the x's.

The computation of η_y may be carried out by adding a few more rows to the form already given in the table on page 212. The work is shown below:

Digitized by Google

x	50–65	65–80	80–95	95–110	110–125	125–140	140–155	155–170
$\bar{y}_k$	33.17	39.50	41.13	44.99	46.40	52.68	52.68	51.64
$\bar{y}_k - \bar{y}$	−9.02	−2.69	−1.06	2.80	4.21	10.49	10.49	9.45
$(\bar{y}_k - \bar{y})^2$	81.36	7.24	1.12	7.84	17.72	110.04	110.04	88.30
$f(\bar{y}_k - \bar{y})^2$	2440.8	477.8	93.0	478.2	372.1	1210.4	1210.4	618.1

$$\Sigma f(\bar{y}_k - \bar{y})^2 = 6900.8 \quad \Sigma f(\bar{y}_k - \bar{y})^2/N = 23.80 \quad \sqrt{\Sigma f(\bar{y}_k - \bar{y})^2/N} = 4.879$$
$$\eta_y = 4.879/\sigma_y = 4.879/9.569 = .501. \quad \eta_y^2 - r^2 = .013$$

A few points in regard to this computation should be noted. The first row, the $\bar{y}_k$, can often be computed almost by inspection; another way is to make use of the fxy row. If any entry in this row is divided by x, which is legitimate because the x of all items of which it is composed is the same, the result is fy, where the y means deviation from the assumed mean. Dividing this by f_x we have the correction to apply to the assumed mean to give $\bar{y}_k$, in class intervals. This is the easiest method for computing $\bar{y}_k$ if the numbers involved are large and the fxy row has already been found.

Measures of linearity of the regression equations. The natural measure of the closeness with which a regression equation fits the line of the means, must in some way be based on the differences between the y on the line and the corresponding column-mean. For the line, $y - \bar{y} = m_x(x - \bar{x})$ or $Y = m_x X$. The difference wanted is then $\bar{y}_k - y = \bar{y}_k - [\bar{y} + m_x(x - \bar{x})] = \bar{y}_k - \bar{y} - m_x(x - \bar{x})$. Let us denote $\bar{y}_k - \bar{y}$ by Y_k. The difference is then $Y_k - m_x X$. The root-mean-square of these differences is analogous to the standard deviation, and the ratio of this expression to the original standard deviation of the y's, seems a natural measure to use for closeness of fit. The formula is

$$\frac{\sqrt{\Sigma f_x (Y_k - m_x X)^2}}{N\sigma_y} = M$$

which may be simplified as follows: Let us deal with $M^2\sigma_y^2$ first. Expanding, we have

$$\frac{\Sigma f_x Y_k^2}{N} - 2m_x \frac{\Sigma f_x Y_k X}{N} + m_x^2 \frac{\Sigma f_x X^2}{N} = M^2\sigma_y^2$$

Digitized by Google

But the first summation is $\eta_y^2 \sigma_y^2$, the second exactly p, and the third exactly σ_x^2. We have, then, since $m_x = p/\sigma_x^2$,

$$\eta^2\sigma_y^2 - 2p \times p/\sigma_x^2 + p^2/\sigma_x^2 = M^2\sigma_y^2$$

and further, $$M^2 = \eta_y^2 - p^2/\sigma_x^2\sigma_y^2 = \eta_y^2 - r^2$$

According to Blakeman,[1] a rough criterion of linearity is that M^2 shall be less than $11.5/N$; this test should be applied if the lines of regression do not seem to be linear. If M^2 turns out to be greater than $11.5/N$, other methods should be substituted for the linear regression equations, and the correlation ratio should be used rather than the coefficient.

As M^2 is positive, it follows that η is larger than r. But as η is the ratio of the root-mean-square of the grouped deviations to the root-mean-square of the ungrouped, it can not exceed unity. Unlike r, it is always positive. A more formal proof that it must be less than unity may be given, as follows: The correlation ratio will be less than one, if, using the notation of the preceding page, $\Sigma f Y_k^2 < \Sigma f Y^2$. Consider the contributions made to each of these sums by the items in a single column. The column contributes to the first sum f_x items, each equal to the square of the distance from the column-mean to the general mean. The column contributes to the second sum the same number of items, in general not all the same. If d is the deviation of any item from the column-mean, we have $Y = Y_k + d$. Squaring, we have $\Sigma Y^2 = \Sigma Y_k^2 + 2\Sigma Y_k \times d + \Sigma d^2$. Now Y_k is constant for all the items in a column, so that we have

$$\Sigma Y^2 = f_x Y_k^2 + 2Y_k \Sigma d + \Sigma d^2$$

Now $\Sigma d = 0$, as the d's were taken about the column-mean, and all items in the column are summed. We therefore have $f_x Y_k^2 < \Sigma Y^2$, unless all the d's are zero, which is an exceptional case statistically; in that case the two expressions are equal. For each column, we have an inequality like that stated, or an equation; adding them all up, we have $\Sigma f Y_k^2 < \Sigma f Y^2$, unless in every column all the d's are zero. As this last case is impossible unless the division of the table is carried beyond what the number

[1] *Biometrika*, vol. IV, 1905, p. 332.

Digitized by Google

of items justifies, we may say that the correlation ratio is always less than unity, for any two-way frequency table. It is of course to be noted that the process of finding η is meaningless unless such a table has been formed. The *coefficient* of correlation is sometimes computed without such a table, but its validity should in that case be viewed as open to question, as this check for linearity cannot be applied.

There are two limitations of the correlation ratio which should be borne in mind. If the number of items is small compared to the number of divisions of the table, the ratio will be unduly large. To some extent in such cases, the ratio is a function of the closeness of division of the table. This point is not significant with tables based on over 100 items, say, divided into six or seven divisions each way, or for tables based on more items, and a proportionate (or less) increase in the number of the divisions. The second point is that accidental fluctuations of the column-means may result in fairly large values of the numerator of the correlation ratio, even when there is no real correlation. Correlation is negligible, therefore, unless the correlation ratio exceeds .30 or .35.

The index of correlation. It was shown on page 206 of this chapter that the sum of the squares of the deviations of the actual ordinates from those given by the regression equation is equal to $N\sigma_y^2(1-r^2)$. This expression is also equal to N times the square of the standard deviation of estimate, that is to say, to NS_y^2, if we introduce the symbol S_y to denote standard deviation of estimate, i.e., root-mean-square of the differences between actual ordinates and those given by the regression formula. We have then

$$NS_y^2 = N\sigma_y^2(1-r^2),$$

or, solving for r,

$$r^2 = 1 - S_y^2/\sigma_y^2.$$

This formula could be used for computing the coefficient of correlation, or as a check.

In order to provide a measure of correlation analogous to the coefficient of correlation for a set of paired values in which some

Digitized by Google

definite non-linear equation for y in terms of x has been found, Professor F. C. Mills has suggested using this formula to give what he terms an *index of correlation.*[1] S_y will mean of course the square root of the mean of the squares of the differences between the actual ordinates and those given by the non-linear equation.

For any particular equation, such as the parabola or exponential curve, it is possible to express the index of correlation in terms of the constants of the equation and avoid the separate computation of S_y.

Comparison of rank lists, and the rank-difference coefficient. In some instances, especially when the number of cases is not large enough to make a correlation table possible, or the information is such that the absolute value of the variable is not known, but it is possible to rank each item as greater or less in certain characteristics than every other item, it is desirable to determine correlation by the use of the rank lists. This may be done roughly, merely by arranging the items as a rank list according to one variable, and writing beside each item its rank according to the other variable. If the high ranks on the second basis come on the whole near the top, we have high correlation; if they come near the bottom, negative correlation; if they are scattered indiscriminately, no correlation. A glance at such a list gives a rough idea of the facts, but for more accurate work it is necessary to compute a modified form of the coefficient of correlation.

We need for this purpose a formula for the coefficient of correlation which is also useful in certain other problems. If we note for each item the difference of the variables X and Y, it is clearly possible to find the standard deviation of the quantity $X - Y$, and we shall denote this by σ_{x-y}. We have

$$(X - Y)^2 = X^2 - 2XY + Y^2$$

for each item, or summing for all items,

$$\Sigma(X - Y)^2 = \Sigma X^2 - 2\Sigma XY + \Sigma Y^2$$

[1] "The Measurement of Correlation, and the Problem of Estimation," by F. C. Mills, in the *Journal of the American Statistical Association*, Sept., 1924, vol. XIX, N.S. no. 147.

Digitized by Google

Hence

$$N\sigma_{x-y}^2 = N\sigma_x^2 - 2Nr\sigma_x\sigma_y + N\sigma_y^2; \text{ and hence}$$

$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$$

This formula may be used to advantage in cases in which x and y are variables of about the same range of values, and the correlation is high. The computation of σ_{x-y} is carried out by making a new frequency table for $X - Y$, which may be formed readily from the original two-way table by adding frequencies along the diagonals. The correction term for $X - Y$ is $c - d$, to bring results from assumed to true mean.

We come now to the application of this formula to the ranks. In this case, both x and y take the succession of values 1, 2, 3, 4, etc. and $\bar{x}$ is the mean value of the first N natural numbers, i.e., $N+1/2$. We have also $\sigma_x^2 = \Sigma N^2/N - (N + 1/2)^2 = \sigma_y^2$. Now it is shown in algebra, or in the following note, that

$$\Sigma N^2 = (2N + 1)(N + 1) \times N/6 \text{ (cf. page 165).}$$

We have, therefore,

$$\sigma_x^2 = (N + 1)\ (2N + 1)/6 - [(N + 1)/2]^2 = (N^2 - 1)/12$$

Therefore, $R = (2\sigma_x^2 - \sigma_{x-y}^2)/2\sigma_x^2 = 1 - 6\Sigma(R_1 - R_2)^2/N(N^2 - 1)$, where R_1 denotes the rank of an item in the first characteristic, and R_2 in the second. It may be shown that by the operation of chance alone the value of $\Sigma(R_1 - R_2)^2$ would be $N(N^2 - 1)/6$; so that in that case R would be zero. If the correlation as measured by ranks were perfect, R_1 would be equal to R_2 for every item, and hence R would be unity. It may be shown that if each distribution is normal, i.e., if the frequency polygon for each characteristic conforms closely to the normal curve, to be discussed later, then this R computed from the rank differences bears a definite relation to the r computed by formula previously given, which is known sometimes as the product-moment formula. This relation is that $r = 2 \sin (\pi R/6)$. That is to say, multiply 30° by R and take twice the sine of that angle. The use of this transformation is not recommended in numerical work, however, as it only

Digitized by Google

expresses what would be true of a theoretically perfect distribution, and does not give exactly what one will find in a particular case.

As the rank difference method, without the determination of standard deviations, correlation ratio, and construction of a correlation table, does not permit a really adequate discussion of the case, it should be used only for cases in which no information except the ranks is available or when the number of items is less than 30.

Note: Proof that $\Sigma N^2 = N(N+1)(2N+1)/6$:

$(N+1)^3 - N^3 = N^3 + 3N^2 + 3N + 1 - N^3 = 3N^2 + 3N + 1$

$N^3 - (N-1)^3 = 3(N-1)^2 + 3(N-1) + 1$, by replacing N by $N-1$ above.

$(N-1)^3 - (N-2)^3 = 3(N-2)^2 + 3(N-2) + 1$, by replacing N by $N-2$ above.

$(N-2)^3 - (N-3)^3 = 3(N-3)^2 + 3(N-3) + 1$, and so forth;

....

$2^3 - 1^3 =$ $3 \cdot 1 + 3 \cdot 1 + 1$

Adding, we have $(N+1)^3 - 1^3 = 3\Sigma N^2 + 3\Sigma N + N$, the summations running from 1 to N in each case.

Transposing, expanding $(N+1)^3$, and simplifying, we have: $3\Sigma N^2 = N^3 + 3N^2 + 2N - 3\Sigma N$. Now the sum of the first N natural numbers is $N(N+1)/2$. [For we have $S = 1 + 2 + 3 + \ldots + N = N + (N-1) + (N-2) + \ldots + 1$, and adding the two first terms in each of these ways of writing the sum, we have $N+1$; we have the same thing for the second term, etc. In all we have $2S = N(N+1)$.]

So $3\Sigma N^2 = N^3 + 3N^2 + 2N - 3N(N+1)/2$

or $6\Sigma N^2 = 2N^3 + 6N^2 + 4N - 3N^2 - 3N = 2N^3 + 3N^2 + N$

$6\Sigma N^2 = N(N+1)(2N+1)$ and $\Sigma N^2 = N(N+1)(2N+1)/6$

Example showing computation of rank-difference coefficient of correlation. The following table shows in detail the method for computing the rank-difference coefficient of correlation in the case of the 26 school girls whose records in a test and ages were given at

Digitized by Google

the beginning of this chapter. The ranking in the second and third columns is easily written down with the aid of the scatter diagram.

Girl (by number)	Rank in Age — Youngest First	Rank in Test	Difference in Ranks	Squared Diff. of Ranks
1	3	1.5	1.5	2.25
2	6.5	9	2.5	6.25
3	8	5	3	9
4	23	26	3	9
5	22	20	2	4
6	16	8	8	64
7	2	6.5	4.5	20.25
8	18.5	16	2.5	6.25
9	18.5	6.5	12	144
10	16	24	8	64
11	1	1.5	.5	.25
12	20	22.5	2.5	6.25
13	10.5	17.5	7	49
14	24	10.5	13.5	182.85
15	10.5	12.5	2	4
16	13	14	1	1
17	13	21	8	64
18	4	15	11	121
19	9	19	10	100
20	21	10.5	10.5	110.25
21	6.5	3	3.5	12.25
22	16	12.5	3.5	12.25
23	5	4	1	1
24	13	17.5	4.5	20.25
25	25.5	22.5	3	9
26	25.5	25	.5	.25
Total.....				1022

$$\text{Hence } R = 1 - \frac{6\Sigma(R_1 - R_2)^2}{N(N^2-1)} = 1 - \frac{6\times 1022}{26\times 25\times 27} = .65$$

It may be noted that in this case the value of the product-moment coefficient of correlation r is .62, and that the supposed value of r obtained by use of the formula $r = 2 \sin (\pi R/6)$ is .67. In this case, therefore, the uncorrected value of R is nearer to the true r than the corrected value is.

Digitized by Google

Summary of Chapter X. An important statistical problem is to determine and express the extent to which the possession of more or less of one characteristic (x) by any individual in a group implies the possession of more or less of another characteristic (y). More briefly, the problem is to determine the relationship between two variable quantities x and y in a group of individuals.

The relationship may be shown:

(1) By a scatter diagram, that is to say, a plot with one point for each individual, the horizontal distance being proportional to the measure of one characteristic and the vertical distance to the other.

(2) By a three by three table, showing how many persons are below average, average, and above average in each characteristic.

(3) By finding the average value of the second characteristic for each of a series of selected values (or ranges of values) of the first characteristic.

(4) By constructing a two-way frequency table.

(5) By stating certain key proportions deduced from the frequency table.

(6) By constructing a three dimensional model showing the frequency corresponding to each value-range of x and y as a column or pole perpendicular to the plane of x and y.

(7) By contour lines representing this surface.

The two-way table portrays relationships which are less uniform and exact than those discussed in the exact sciences, and is an essential means of investigation in the social and natural sciences where it is not possible to conduct carefully controlled experiments in which variations due to extraneous influences are reduced to minor importance.

When the method of averages, (3) above, is used, it is often advisable to fit an equation to these averages by the methods of Chapter VIII.

Analysis of the formula for this equation shows that the number which expresses the essence of the relationship is

Digitized by Google

$$\Sigma f(x-\bar{x})(y-\bar{y})/N\sigma_x\sigma_y$$

This number is called the *coefficient of correlation.*

The coefficient of correlation always lies between -1 and $+1$. If numerically less than .2, it is negligible; if between .2 and .4, the correspondence between x and y exists, but is low; if r is between .4 and .6, there is moderate correspondence; if r is higher than .6, the correspondence is significant, the degree of significance depending on the type of problem, number of cases considered, homogeneity of data, and other factors.

The computation of the coefficient of correlation requires the computation of the average product-sum, $\Sigma f(x-\bar{x})(y-\bar{y})/N$, in addition to the means and standard deviations of x and y, whose computation has already been discussed. This average product-sum can be conveniently computed from the individual pairs of values of x and y, if there are not over 35 or 40 pairs, and from the two-way frequency table, if there are more pairs than that.

An important application of the coefficient of correlation is in determining the average length of time by which the fluctuations of one historical series precede or follow the fluctuations of a second historical series.

In the case of non-linearity of the regression equations, correlation may be measured by the correlation ratio, defined as the root-mean-square deviation of the column (or row) means as compared with the general mean, divided by the general standard deviation. In symbols,

$$\eta_y, \text{ correlation ratio, } = \sqrt{\frac{\Sigma f(\bar{y}_k-\bar{y})^2}{\Sigma f(y-\bar{y})^2}}$$

If $\eta_y^2 - r^2$ is greater than $11.5/N$, the regression of y on x is not nearly enough linear to justify the use of the linear regression equation and the coefficient of correlation.

The rank-difference coefficient of correlation approximates the coefficient of correlation, when the number of items is small, sufficiently well so that it may be used to advantage, especially if only the ranking of the items is known. The formula is

$$R \text{ (rank-difference coefficient of correlation)} = 1 - \frac{6\Sigma(R_1-R_2)^2}{N(N^2-1)}$$

Digitized by Google

CHAPTER XI

ELEMENTARY IDEAS IN REGARD TO SAMPLING

Sampling and probability. Up to the present point in our studies we have regarded statistics almost entirely as a means of describing in brief form a complicated mass of quantitative facts. We discovered that the production of lumber in the United States has been decreasing at such and such a rate; we determined for a particular sample of ears of corn the mean length, the median length, the quartile deviation of the lengths, the mean deviation, the standard deviation, etc. The example considered in Chapter X, to be sure, in which it was found that the general business index number lagged five months behind the stock index number, suggested on its face the possibility of using the facts found as a basis for predicting the future. But even here certain doubts may well suggest themselves. In the first place, is the fact that a certain relationship held true in the past a sound reason for expecting it to hold true in the future? Secondly, how *closely* are we to expect a reproduction of past history by the unfolding events of the future? That is to say, can we count on five months as the interval rather than perhaps four or six? Thirdly, is a prediction of this sort equally entitled to confidence regardless of the details of the construction of the index numbers? If not, how does it vary?

Such questions as these arise in the use of many statistical results. The thing that is really valuable is not the result in itself, but the general law that may be inferred from it. It may be valuable, for instance, to know that ears of corn raised under certain conditions may be expected to have the lengths indicated by the sample measured. We may thus be able to say, by comparing two samples selected from ears raised under two different sets of conditions, that one plan of cultivating the crop or selecting the seed is better than another. But before any such statement can be made, we must have some principle on which to judge whether the sample is a fair one, and also whether the differences observed are

Digitized by Google

really significant. These principles are discussed in what is known sometimes as the theory of random sampling, a part of the theory of probability.

Relation of sampling to ordinary thought and to scientific induction. Before we take up in detail the discussion of this subject, it is well to note that the type of argument discussed is not peculiar to statistics, but is really involved in all inductive thought. What is peculiar to statistics is the fact that the laws found or the statements made are quantitative and definite, and the methods for determining their soundness can therefore be made more exact and be expressed in more mathematical form than in the case of the judgments of common life. These latter judgments, however, are also based on the careful consideration of a sample, the formation of a tentative conclusion, and the testing of that conclusion by comparison with what other people think and by seeing if it holds true in other samples. Some of our most generally accepted truisms were once such tentative conclusions, but are now so thoroughly tested that they are called axioms or proverbs. Even such a statement as that "All men are mortal" does not derive its authority from an actual examination of the histories of all men to see if they have actually died. The real basis is that an examination of the histories of a large number of men suggested to some one that what was true of all men of a certain group — those born before a certain date, say — was true of all men; and this hypothesis has been confirmed so often, and further experience and scientific study have shown such great similarities between the men who have died and other men, that we feel convinced that the sample, consisting of all men of the past, is an adequate basis for a statement concerning all men.

As a second example, consider the statement "Honesty is the best policy." Presumably this was originally made by somebody who had reflected on a number of instances of honesty and dishonesty that had come to his attention, and had come to this conclusion as applicable to these instances and hence to all instances. The "hence" is of course not logically justified; but in ordinary life every one reasons so. What we should say in such cases is, "I sub-

Digitized by Google

mit as an hypothesis that this is always true." Then the gradual comparison of this conclusion with other instances would lead to its general acceptance or rejection. In this case, the verdict of mankind is favorable, and the statement is accepted by most people as true.

Every individual draws up hypotheses for his own guidance in dealing with the problems which confront him. He decides that he is not able to do well in mathematics, for instance. What is the basis for this conclusion? Very often merely that in certain mathematics classes which he has attended he has not done well. He is judging on the basis of a small sample, with no allowance for possible differences if other conditions were encountered — different subject-matter, different teacher, different atmosphere, different attitude on his part. He assumes that mathematics means only the kind of mathematics he studied; that the conditions under which he labored were typical of the conditions found elsewhere; in short that the sample he considered was adequate for forming a judgment. If he considers this judgment merely tentative — i.e., an hypothesis — his idea is reasonable; the soundness of the hypothesis depends on the adequacy and representative character of the sample. So in general we form our opinions of people and affairs, and are known as "opinionated" if our statements are presented as oracles when as a matter of fact they are merely hypotheses, often with slight foundation.

The opposite error in regard to unstatistical opinions is often made by those who have acquired some knowledge of more satisfactory methods. They often think that individual judgments made on apparently very inadequate data and indefinite analysis, are worthless. It will on the contrary often be found that careful statistical analysis merely confirms these judgments, which are more carefully formed than is indicated by the data produced to support them. They have perhaps been formulated long before in the individual's mind as hypotheses, and have been tested by their agreement or disagreement with a long series of particular instances. The individual may not be able to recall many of these instances, or even to say how long the hypothesis has been under consideration; but if he is fair-minded, able and willing to take into

Digitized by Google

due account cases which seem to contradict as well as cases which seem to favor the hypothesis, his opinion has value. It often happens, however, that two apparently fair-minded individuals have diametrically opposite opinions. The difference may be due to differences in the cases which have come to the attention of each individual, or to differences in personal bias. The only way to arrive at the truth in such cases is to conduct an impartial statistical investigation, either by collecting opinions from a large number of fair-minded and well-informed individuals, or, better, by a direct examination of the facts.

The field of education furnishes many good illustrations of the points of the preceding paragraph. Education has been one of the major activities of mankind for thousands of years, but only recently have exact methods been applied to measure the relative effectiveness of different methods of instruction. It would be absurd to infer from this last fact that there is no value in the opinions of the fairest minded educators in regard to such questions as the number of pupils one teacher can handle satisfactorily. As there is not close agreement on this question, statistical studies are necessary for a final answer, but opinions provide a basis for a preliminary answer. On the other hand, the vexed question of the value of the study of Latin is not much illuminated by a collection of opinions, because of the wide divergence of opinion among those who have the best basis for judgment and the emotional atmosphere in which the question has become enmeshed.

The object of statistical methods in general is to provide a means more reliable than mere opinion for reaching conclusions on questions in which the data are quantitative. In particular, the question which now concerns us is the number of cases which must be examined before a conclusion is justified, and the degree of reliability which inheres in any conclusion reached as the result of a statistical investigation.

Fundamental points in the theory of sampling. The last sentence of the preceding paragraph states the aim of the theory of sampling in a way to bring out its bearing on logical thought in general. Unfortunately, before we can even begin to develop the theory, we

Digitized by Google

must restate the problem and develop a new terminology. We must use certain concepts in the mathematical theory of probability, which are inherently difficult to comprehend because it is impossible, or almost so, to define probability without using some ideas which are really based on the definition.

The difficulty will be apparent if we scrutinize again part of the sentence at the end of the preceding section, "the number of cases which must be examined before a conclusion is justified." What do we require to "justify" a conclusion? From the previous discussion, it is clear that when we are attempting to state a general principle which will be true in the future, we can never be absolutely certain that we are right. There must, then, be a thing which we may call degree of certainty. For the present, until we can make the idea more precise for certain cases, let us assume that there are several recognizable degrees of certainty, and that it is possible to know that a particular conclusion has a particular degree of certainty. The first part of the problem of sampling may then be stated, How large a sample (how many cases) must be examined to secure any required degree of certainty in the result?

In a similar way, let us scrutinize the second phrase in the last sentence of the preceding section, "the degree of reliability which inheres in any conclusion." If a conclusion is perfectly reliable, the outcome of a future event will be exactly forecast by the conclusion. "Degree of reliability" implies that the forecast will not always be absolutely correct. That is to say, we may in some sense "expect" an error if we apply the conclusion from one set of circumstances to another set of circumstances which seems to be about the same as the first. The second part of the problem of sampling may then be stated, What variation in the results should be expected if another group of cases were selected in the same way that the original group was selected?

It is to be noted that both questions of the theory of sampling assume that the number of cases which might be examined is larger than the number actually examined. The theory is therefore not concerned with statistical description, where summaries are made of an entire group, until those results are used to predict the facts for other similar groups or for the same group at a future time.

Digitized by Google

In order to fix our ideas in regard to the second part of the problem of sampling, let us think how we should deal with the following problem: It is found that of a group of 100 men of age 25, 60 per cent are over 5 feet 6 inches tall. What variation in this percentage may be expected if another group of 100 men is selected in the same way that the first group was selected?

Now it is found that this problem cannot be dealt with directly as readily as a somewhat similar problem, namely, if 60 per cent of all men of 25 are 5 feet 6 inches tall, what variation in this percentage may be expected if successive groups of 100 men of that age are examined? In this form, the problem is clearly equivalent to that of the percentage of white balls in a set of 100 balls drawn from an urn in which a certain proportion (p), in this case three fifths, of the balls are white and the remainder black. "Common sense," so-called, is probably sufficient to tell most of us that if 100 balls are drawn from the urn, the composition remaining always as stated, three fifths of the balls drawn, or 60, would in some sense be the most natural number of white balls. But obviously it would be possible, theoretically, that all the balls should be white, or none of them, although by common sense we should know that such a result as either of these is "unlikely." But the further question as to the relative probability of 59, 60, 61, etc., can be answered only by an appeal to the theory of probability.

As the subject may not be familar to all readers, it is necessary to give a few definitions. If an event may happen in F ways, and fail to happen in U ways, each of these ways being equally likely to occur, the probability of the event is then said to be $F/(F+U)$. The phrase "equally likely to occur" is not perhaps self-explanatory, as it involves the idea of likelihood or probability, which is being defined: "Equally likely" means that as far as our knowledge goes, there is no reason for thinking that there is any difference between the ways, beyond the difference which establishes their identity. The balls, for instance, are to be of equal size, and smoothness, and weight, and are to differ in nothing except color.

If p is the probability of one event, and p' of another independent event, the probability that both will occur is the product of the probabilities, pp'. For if the first event can happen in F ways,

Digitized by Google

and fail in U ways, and the second event can happen in F' ways and fail in U' ways, then as each way for the first event can happen with each way for the second event, both events can happen in FF' ways, and the total number of ways of happening or not happening for the two events is $(F + U)(F' + U')$, and the ratio favorable to total is $FF'/(F + U)(F' + U')$ which is obviously $= pp'$.

For instance, if two balls are drawn from the above mentioned urn, the first ball being replaced and shaken up with the others before the second is drawn, the probability of drawing a white ball the first time is three fifths, and in three fifths of these cases a white ball will be drawn the second time, so that the probability of drawing 2 white balls is $3/5 \times 3/5 = 9/25$. Similarly the probability of drawing a white and then a black ball is $3/5 \times 2/5 = 6/25$, and of drawing a black and then a white ball $2/5 \times 3/5 = 6/25$, and finally of drawing two black balls is $2/5 \times 2/5 = 4/25$. It will be noted that as always, the sum of the probabilities of all possible events is unity. This summation is a particular application of another principle, that the probability that two independent and mutually exclusive events will both occur is the sum of the separate probabilities.

In this case, for two independent drawings, we note the following probabilities: 2 white, 9/25, 1 white, 1 black, 12/25; 2 black 4/25. For three balls, we easily find the following probabilities:

WWW, $(3/5)^3 = 27/125$; WWB, $(3/5)^2 (2/5) = 18/125$ [WBW and BWW the same]; WBB, $3/5(2/5)^2 = 12/125$ [BWB and BBW the same]; BBB, $(2/5)^3 = 8/125$. If the order of occurrence is not considered, the probability of 2 white and one black is $3(3/5)^2 (2/5)^2$ and of one white and 2 black, $3(2/5)^2 (3/5)$. Let us now compare these various probabilities with the values of successive terms in the expansion of $[3/5 + 2/5]^3$, i.e., with

$$(3/5)^3 + 3(3/5)^2 (2/5) + 3(3/5)\ (2/5)^2 + (2/5)^3.$$

We see that the terms are respectively the probabilities of 3 white, 2 white and 1 black, 1 white and 2 black, and 3 black.

As we shall show algebraically in the next chapter, the expansion of $(p + q)^n$ always provides a means of finding readily the proba-

Digitized by Google

bility that if n balls are drawn with a probability p of a white ball on each draw, exactly r of them will be white. The required probability is then that term in which the exponent of p is r.

Summary of Chapter XI. Previous chapters showed how to describe masses of quantitative facts in compact form. Before any relations so found can be applied to predict the future or estimate the unknown, we must become familiar with certain modes of thought in what is known as the theory of sampling.

The type of argument used in the theory of sampling is merely a more precise form of that used in all inductive thought, namely, formulation of an hypothesis, test of that hypothesis by comparison with other data and by analysis of related facts, and a tentative decision as to the degree of reliability of the hypothesis.

The fundamental questions in the theory of sampling, are:

(1) How many cases must be examined before we can reach a conclusion of any specified degree of certainty?

(2) If we have formed an hypothesis based on a specified number of cases, what variation in the result may be expected if another group of cases is selected in the same way as the original group was selected?

It is found that these problems cannot be solved by a direct approach, but that it is necessary first to discuss simpler problems of drawing balls from an urn which contains black and white balls in a known proportion. If the probability of drawing a white ball (i.e., the proportion of white balls) is p, and the probability of drawing a black ball is q, then the probabilities if three balls are drawn in succession with replacement are given by the successive terms of the expansion of $(p + q)^3$.

EXERCISES

1. If two dice are thrown, what is the probability that the total number of spots on the topmost sides of the two dice will be 9? What that the number will be 8? What that it will be 8 or 9? [Solve by using the definition of probability and counting up all the favorable and all the unfavorable cases, treating each part of the question as a separate problem.]

2. If an urn contains 25 black balls and 75 white balls, and the ball drawn

Digitized by Google

is replaced after each draw, what is the probability of getting two black and one white ball in three draws? Of getting one black and three white balls in four draws?

3. Read carefully two or three newspaper or magazine articles on non-technical subjects, noting the number of cases in the samples on which the author's statements and conclusions are based. If the size of the samples seems too small to justify the conclusions, try to think how large the samples should be. Note also whether the samples are representative.

4. A ball player whose batting average is .300 comes to bat four times in a certain game. What are the chances of his getting exactly two hits?

Digitized by Google

CHAPTER XII

THE BINOMIAL EXPANSION AND THE NORMAL CURVE

Algebraic proof that probabilities are given by terms of $(p+q)^n$. Let us express algebraically the several probabilities if n balls are drawn, with the constant probability p that the ball drawn is white. The probability that a white ball is drawn every time is clearly $p \times p \times p \ldots$ to n factors, that is to say, p^n. The probability that a white ball is drawn every time except the first is $q \times p^{n-1}$; every time except the second, the same; and so forth. That is to say, there are n times, at each of which a black ball may be drawn, with all the rest white. As the probability of each of these n times is $q \times p^{n-1}$, the total probability of $n-1$ white and one black is $np^{n-1}q$.

Similarly the probability that the first two balls will be black and the others white is $q^2 \times p^{n-2}$; the probability that the first and third will be black and the others white is the same; that the second and third will be black and the others white, the same; etc., that is to say, this probability will be repeated as many different times as two things can be selected in a different way from n things. We can make the first of these two selections in n ways, and after having done this can make the remaining selection in $n-1$ ways, making a total of $n(n-1)$ ways. These ways are not all different, however, as for our purposes the combination first and third, for instance, is the same as third and first. Each way is given twice; we therefore obtain the number of different ways by dividing by 2. The total probability of $n-2$ white and 2 black is therefore $n(n-1)q^2p^{n-2}/2$.

These three terms are exactly those of the binomial expansion, and a continuation of the same argument, involving rather more use of the theory of permutations and combinations, shows that every probability is given exactly by the corresponding term of the expansion. Consider the probability that the first r balls are white and the remainder black. This is clearly p^rq^{n-r}. Now this

Digitized by Google

same probability holds for *any* r balls being white. The coefficient is therefore the number of different ways we can select r balls from n balls. As before, we can select the first in n ways, the second in $n-1$ ways, the third in $n-2$ ways, and finally the rth in $n-r+1$ ways. We have then $n(n-1)(n-2)\ldots(n-r+1)$ ways, not all different. The question now is, how many times is any given combination really repeated? The first position in any specified selection may have been filled in r different ways; the second in $r-1$, and so forth. All told, any given combination is therefore repeated $r(r-1)(r-2)\ldots 2\cdot 1$ times, and the number of different combinations is really then

$$\frac{n(n-1)\ (n-2)\ (n-3)\ldots(n-r+1)}{1\cdot 2\cdot 3\ldots(r-1)r}p^r q^{n-r}$$

This is the probability that exactly r balls will be white if n are drawn, the probability of drawing a white ball on each draw being always p. It is also the general term of the binomial expansion, and therefore it is usual to discuss many of the problems of sampling by the use of this expansion. It is advisable to get this connection carefully in mind before going on to the algebraic development.

Consider the following application of the formula: Suppose a large number of school children are required to do 20 tasks, in each of which the probability of success is three fifths. What is the probability that a child will do correctly exactly 4 [or any number] of these tasks correctly? Here $n = 20$, $p = 3/5$, $q = 2/5$. Carrying out the arithmetic, we find the following probabilities:

Correct:	20	19	18	17	16	15	14	13	12	11	10
Probability:	.00004	.0005	.0031	.0124	.0350	.0746	.1244	.1659	.1797	.1597	.1171

Correct:	9	8	7	6	5	4	3	2	1	0
Probability:	.0710	.0355	.0146	.0049	.0013	.0003	....	....	....	....

Now if 10,000 children are required to perform these 20 tasks, the number who will do exactly 12 correctly will be $.1797 \times 10{,}000$, as for each of the children the probability of doing that number is .1797, and therefore the frequency distribution of children under these conditions is obtained by removing the decimal point in the above table (except for the entry under 20).

Digitized by Google

As this is the first frequency distribution with a theoretical basis, which we have dealt with, it is worth while to pause to examine it. We note a mode at 12, and a lack of symmetry about that mode. If we plot the frequency polygon, we see that the ordinate for a number of problems just less than 12 is lower than the corresponding ordinate at an equal distance on the other side of 12; but that the ordinate at 8 is higher than that at 16, and similarly for the other ordinates at the end. The sum of ordinates for number of problems greater than 12 is .4159, and for number of problems less than 12 is .4044, a further result of the lack of symmetry.

Position of mode of binomial distribution. In the preceding example, the highest point of the frequency polygon happened to occur for the number correct $r = np$. As this is always nearly true, it is desirable to discuss the situation generally, that is to say, algebraically. What we want to do is to pick out the largest term of the binomial expansion. We shall do this by writing down the formulas for three consecutive terms, and stating the conditions that the second term shall be either (1) equal to the first, and greater than the third, or (2) greater than the first and also greater than the second. It will also be shown that in either case, all other terms are smaller than the second of the three, if the conditions found are satisfied.

Three consecutive terms are those in which the exponents of p are respectively $r + 1$, r, and $r - 1$. Making use of the formula for the general term, the three consecutive terms are:

$$\frac{n(n-1)(n-2)(n-3)\dots(n-r)}{1\cdot 2\cdot 3\dots(r+1)}p^{r+1}q^{n-r-1}$$

$$+\frac{n(n-1)\dots(n-r+1)}{1\cdot 2\cdot 3\dots r}p^{r}q^{n-r}$$

$$+\frac{n(n-1)\dots(n-r+2)}{1\cdot 2\cdot 3\dots(r-1)}p^{r-1}q^{n-r+1}$$

The second term will be equal to the first and greater than the third if

Digitized by Google

$$(n-r)p = (r+1)q \text{ and } (n-r+1)p > rq$$

i.e., $$np - q = r\,(p+q) = r \text{ and } np + p > r(q+p) = r$$

since $p+q=1$. The second condition adds nothing to the requirements, for if $np = q+r$, according to the first condition, then $np+p = p+q+r = r+1$, which is obviously $>r$. But r can be exactly equal to $np-q$ only if the latter expression is an integer, which is true only for certain values of n and p. There will therefore be two equal maximum terms in the binomial expansion only in certain cases, namely when $np-q$ is an integer, and the two equal maximum terms in such cases will be those in which the exponent of p is $np-q$ or $np-q+1$.

The usual case is that of a single definite mode, in which one term is larger than the term before and also larger than the term after, and the following relations are therefore true:

$$(n-r)p < (r+1)q \text{ and } (n-r+1)p > rq$$

Hence we find $np - q < rp + rq = r$ and $np + p > rp + rq = r$

That is to say, the exponent of p in the maximum term is greater than $np-q$ and less than $np+p$. As this exponent must of course be an integer, it is definitely located by these inequalities, as the difference between the limits is $1(np+p-[np-q] = p+q = 1)$.

In order to include the possibility of two equal maximum terms in one general statement, we may summarize as follows:

The maximum term of the binomial expansion is that for which r, the exponent of p, is the integer which falls between $np-q$ and $np+p$, except that, if $np-q$ is an integer, the terms in which $r = np-q$ and the term in which $r = np+p$ will be equal and each larger than any other term.

We have not yet really shown that the maximum terms selected as above are really larger than *every* other term, but only that they are larger than adjacent terms. The fact is that the magnitude of the terms in the binomial expansion $(p+q)^n$ rises regularly from the first term to the maximum, and then decreases. This may easily be seen, by considering the factors by which we would multiply to get from the terms written out above to earlier or later terms. For instance, to get the term in which the exponent of p

Digitized by Google

is $r+2$, we must multiply the first term above by $(n-r-1)p$ and divide by $q(r+2)$. This will result in a decrease if

$$(n-r-1)p < q(r+2)$$

i.e., if $$np - p < pr + qr + 2q$$

i.e., if $$np - q - 1 < r$$

But, as shown above, r is greater than $np-q$, and therefore is certainly greater than $np-q-1$, and hence the term in which the exponent of p is $r+2$ is certainly smaller than the term in which the exponent of p is $r+1$.

It is desirable to restate these algebraic conclusions in the terminology of probability. As r is the number of successes out of n possible successes, we have really shown above that the number of successes which will occur most frequently in a large number of samples each of size n, is the integer between $np-q$ and $np+p$. Or in other words, the most probable number of successes never differs from np by as much as unity.

It may be noted that the location of the mode, as explained above, materially facilitates the computation of probabilities by means of the binomial theorem. It will be found that a short process of computation, which can easily be carried through very accurately, is as follows:

(1) Determine the exponent of p in the maximum term.

(2) Compute the maximum term. If tables of $n!$, "factorial n" (the product of all integers up to and including n), are available, this computation can be simplified by using the formula for the rth term in the following form, which may be verified by the student:

$$\frac{n!}{r!(n-r)!}p^r q^{n-r}$$

(3) Determine next the terms adjacent to the maximum term, by multiplying and dividing the result of (2) by the appropriate factors. These factors can be readily determined by inspection of the formulas of the three middle terms given

Digitized by Google

above. The factors are simple numbers in most cases and lend themselves to rapid arithmetic.

(4) Similarly find the terms adjacent to those found in (3), and so on till all terms large enough to be significant have been found.

This method is much superior to beginning with the first or last terms and working through, in that accuracy can only be secured in that method by carrying the work to a large number of decimal places.

Effect of increasing n. Two other expansions by the binomial theorem will now be taken up to show the effect of increasing n. If we take $p = q = 1/2$, we have the following probabilities of number of successes, for $n = 4$, 9, and 16, respectively.

No. of successes	0	1	2	3	4	5	6	7	8
$n = 4$	.0625	.25	.375	.25	.0625				
$n = 9$	.00195	.0176	.0703	.1641	.2461	.2461	.1641	.0703	.0176
$n = 16$	.000015	.00024	.0018	.0085	.0278	.0668	.1222	.1744	.1962

No. of successes	9	10	11	12	13	14	15	16
$n = 9$	.00195							
$n = 16$	.1744	.1222	.0668	.0278	.0085	.0018	.00024	.000015

We may note the following comparisons:

(1) The probability of the most probable number of successes decreases as n increases.

(2) In general, the probabilities of numbers of successes within any specified interval from the mode, decrease as n increases.

(3) The probabilities of complete success or complete failure decrease as n increases; in fact, the ends of the distribution as a whole get smaller.

As the reader may have thought, these distributions can be fully analyzed and compared only after we have found their means and standard deviations. It is therefore desirable to derive algebraically at this time formulas for these two quantities in the case of the general binomial expansion.

Digitized by Google

Proof that mean = cnp and $\sigma = c\sqrt{npq}$ for binomial distribution. In order to make our ideas definite, let us think of the successive terms of the binomial expansion, starting with that corresponding to no successes, as the ordinates of a frequency polygon. Let us call the common distance between these ordinates, c; and take the position of the "no successes" ordinate as the assumed mean. The computation of the mean and standard deviation in the usual way is then shown in the following table:

Deviation	Frequency (ordinate)	Deviation × Frequency	Deviation² × Frequency
0	q^n	0	0
c	npq^{n-1}	$cnpq^{n-1}$	c^2npq^{n-1}
$2c$	$\frac{n(n-1)p^2q^{n-2}}{2}$	$\frac{2cn(n-1)p^2q^{n-2}}{2}$	$2c^2n(n-1)p^2q^{n-2}$
.	.	.	.
rc	$\frac{n(n-1)\dots(n-r+1)p^rq^{n-r}}{1\times2\times3\times\dots\times r}$	$\frac{cn(n-1)\dots(n-r+1)p^rq^{n-r}}{1\times2\times3\times\dots(r-1)}$	$\frac{rc^2n(n-1)\dots(n-r+1)p^rq^{n-r}}{1\times2\times3\times\dots\times(r-1)}$
.	.	.	.
nc	p^n	cnp^n	$c^2n^2p^n$

To determine the true mean, we must, according to the usual procedure, find the sum of the terms in the third column. Taking out the factors common to all terms, we find that this sum is

$$cnp[q^{n-1} + (n-1)\,pq^{n-2} + \dots + \frac{(n-1)\dots(n-r+1)}{1\times2\times3\times\dots(r-1)}p^{r-1}q^{n-r} + \dots + p^{n-1}]$$

The expression in brackets is easily seen to be the expansion of $[p+q]^{n-1}$ and is = 1, since $p+q=1$. As the total frequency is 1, the correction to be applied to the assumed mean to give the true mean is therefore cnp. If c is taken to be one, as it often is, the result is np, which is often stated to be the mean of the binomial expansion. Occasionally the mean is measured from a point c units to the left of the "no success" ordinate, and the mean is then of course $c(np+1)$. The mean number of successes must be said to be np; but in certain other interpretations and uses of this expansion, it seems more natural to use cnp, allowing the distance between ordinates to be other than unity.

Digitized by Google

The summation of the last column, for use in finding σ, is somewhat more difficult. Taking out the factors common to all terms, we find that this sum is

$$c^2np[q^{n-1} + 2(n-1)pq^{n-2} + \ldots + \frac{r(n-1)\ldots(n-r+1)}{1\times 2\times 3\times \ldots (r-1)}p^{r-1}q^{n-r} + \ldots np^{n-1}]$$

Now this expression in brackets begins in the same way as the expression in brackets just dealt with, and differs from it only in the value of the coefficients. Suppose then that we write the expression as two parts, one of them the expansion of $(p+q)^{n-1}$ and the other the difference term by term of the two expressions. The sum then is:

$$c^2np[(q+p)^{n-1} + (n-1)pq^{n-2}\ldots + \frac{(r-1)(n-1)\ldots(n-r+1)}{1\times 2\times 3\times \ldots (r-1)}p^{r-1}q^{n-r} + \ldots (n-1)p^{n-1}]$$

Now the latter part of the expression in brackets contains a factor $(n-1)p$. Taking out this factor, we have remaining:

$$q^{n-2} + \ldots \frac{(n-2)\ldots(n-r+1)}{1\times 2\times 3\times \ldots (r-2)}p^{r-2}q^{n-r} + \ldots p^{n-2}$$

which is exactly $(q+p)^{n-2}$, which is equal to unity. The sum of the terms in the last column therefore reduces to $c^2np[1+(n-1)p]$. This is σ'^2, as the total frequency is one. We have, as usual,

$$\sigma^2 = \sigma'^2 - (cnp)^2 = c^2np[1+(n-1)p] - c^2n^2p^2 = c^2np - c^2np^2$$
$$= c^2np(1-p) = c^2npq; \text{ and } \sigma = c\sqrt{npq}$$

This is the most important single formula in this chapter.

Application of the formula $\sigma = c\sqrt{npq}$. Let us use this formula in connection with the arithmetic instances considered earlier in this chapter. Take for instance that of the children assigned 20 tasks. The mean, as $c = 1$, $n = 20$, $p = .6$, $q = .4$, is $20 \times .6 = 12$, which happens in this case to be also exactly the mode, and

Digitized by Google

$\sigma = \sqrt{20 \times .6 \times .4} = \sqrt{4.8} = 2.2-$. As previously suggested, about 2/3 of the cases lie within σ of the mean; the exact figure in this case is 74.7 per cent. If the conditions of the problem had been that the children were assigned 150 tasks apiece, the mean would have been 90 and the standard deviation 6; and as before we would find about 7/10ths of the items within an interval of σ to one side or the other of the mean. For the larger number of tasks, we have thus a spread of 12 corresponding to a spread of 4.4 in the case of a smaller number of tasks. It is to be noted, however, that the spread of 12 is a smaller proportion of the entire range than 4.4 is of the entire range in the other case, the figures being 12/150 = .08 and 44/200 = .22, respectively. Or in general the ratio of the distance $-\sigma$ to $+\sigma$ to the entire range (nc) is $2\sqrt{npq}/n = 2\sqrt{(pq/n)}$. We may say then that the proportionate deviation from the mean decreases as the square root of n. Or to put it another way, as the size of the sample is increased, a larger proportion of the samples will indicate about the same ratio of favorable to total cases as is indicated by the value of p. For $n = 20$, for instance, we found that over one quarter of the samples had number of successes differing by more than two from 12, suggesting values of p of more than .7 or less than .5. For samples of 150, however, the corresponding values of p which are suggested are 96/150 and 84/150, that is to say, .64 and .56. If n is still further increased, the interval will be still further narrowed.

The exact formulation of a statement as to how large n must be in order that this interval may be not greater than a certain assigned width for a certain definite proportion of the samples (not necessarily one fourth, as above), was first made by James Bernouilli in 1709. This formulation makes it logically sound to say that in the limit, as the size of the sample becomes infinite, the proportion of favorable cases is the same as p, the theoretical proportion. This is often given as an alternative definition of probability, and is especially significant in practical problems, as the theoretical proportion of men aged 40 who will die in the next year, or any other useful probability, is not known, but we can observe a large number of cases and note the actual ratio of deaths. Actually the case is not as simple as in the case of games of chance, as the

Digitized by Google

units are not in fact all equal in practice. No group of men whose mortality is being examined is made up of men of equal vitality; and it is unlikely that the group about whom predictions are to be made, is composed of the same proportions of different types of men as the standard group on which the table of mortality was based. In view of this possibility of error, it is incumbent on any one making such a prediction to try to avoid gross errors due to differing composition of the groups, and to recognize that his prediction cannot be exact, stating if possible in some form how large the error of prediction is likely to be.

Verification of the theory of sampling. In view of this use of an actual relative frequency as equivalent to a theoretical probability, it is interesting to consider some experimental verifications of points in the preceding discussion. The following experiment was conducted by the classes in Genetics at Brown University in the year 1922–23. Ten coins were tossed 100 times more or less by each of the 60 members of the classes, and the number of heads in each throw was noted. For each member of the class there was therefore a frequency distribution of 100 items, such as the following:

No. of heads	2	3	4	5	6	7	8	9
No. of tosses	2	18	12	28	17	13	9	1

Now, theoretically, the distribution is that of $[1/2 + 1/2]^{10}$, which is

$$1/1024\,[1 + 10 + 45 + 120 + 210 + 252 + 210 + 120 + 45 + 10 + 1]$$

The comparison of actual and theoretical results is as follows:

No. of heads	0	1	2	3	4	5	6	7	8	9	10
No. of tosses											
Actual	2	62	249	722	1201	1484	1192	722	327	89	10
Theoretical	6	59	265	705	1234	1481	1234	705	265	59	6

The close agreement in general will be noted at once. But the question at once arises, Is the agreement as close as it ought to be? Is a value 327 where 265 is the theoretical value, permissibly close? We can test this by considering the whole 6020 tosses as a sample, and considering what results would follow if such samples

Digitized by Google

were taken a very large number of times. The probability of the event "securing 8 heads" is 45/1024, and the size of the sample is 6020. The mean frequency of 8 heads would then be np, i.e., $6020 \times 45/1024 = 265$, and the standard deviation about this mean would be $\sqrt{npq} = \sqrt{265 \times 979/1024} = 15.9$. The actual deviation of 62 [327 − 265] is therefore just under 4σ, an unusual but far from impossible deviation. The reader may test the other figures in the same way, and will note that the values of σ for the tails of the distribution are relatively larger, and therefore getting 2 or 10 instead of an expected 6 is not so much of a departure from theory as the case just considered, 327 instead of 265.

Many scientists, philosophers, and mathematicians, have been attracted by the task of trying out the theory of probability, and the net result of their experiments has been a general acceptance of the principles and formulas. The two fields in which the application has to date been most thoroughgoing are in the theory of error of physical and astronomical measurements, and in actuarial theory. In the former case, where it leads to the theory of least squares, it has been accepted as the proper basis for the reduction of observations; in the second case, it has become the basis of the whole financial and social structure of life insurance. The necessity for forms of statement which always imply approximation, and a large number of cases, makes the theory unattractive to many mathematicians, accountants, and practical men. But the point of view of this book, and of mathematical statisticians in general, is that these peculiarities are essential in the nature of thought, and must be studied until they seem natural and reasonable.

Other applications of the formula $\sigma = \sqrt{npq}$. The formula $\sigma = \sqrt{npq}$ is one of the most valuable formulas in mathematics. Let us consider some further applications of it, which will shed more light on the topic of the chapter. For instance, a study of the careers of the graduates of three women's colleges showed that in the ages 35–44 the death rate was 4.15 per thousand, as compared with a general rate in the United States registration area of 8.03 per thousand and an insurance experience of 4.83 per thousand. The

Digitized by Google

question which naturally arises in interpreting these figures is, Can this difference be merely the result of chance? That is to say, if other groups of the same size as the college group were selected at random from the other aggregates, would there sometimes be as large a departure from the general rate? We clearly need to know how many college women there were in the group taken; the figure given in the article is that the number "at risk," i.e., the number of women times the number of years each had been exposed to the chance of dying, was 45,240, which is accordingly the size of the sample. Of these 188 actually died, as compared with 222 who would have died at the rate for the insurance experience and 363 at the rate for the registration area. Can chance variation from 222 or 363 give 188? Computing σ from the rate for insurance experience, we find it is 15, and there the actual difference of 34 is over 2σ distant from the mean, i.e., the expected value. Such a deviation is rather unusual, but occurs occasionally, and the difference can only be said to be "possibly" significant. On the other hand, we find that σ for the registration area rate is 19, and the actual difference of 175 is therefore over 9σ, and indicates definitely that such a difference could not have arisen by chance. We may say, then, that the graduates of women's colleges have a lower mortality rate than the general population. We cannot say without other evidence that this is the result of their college training, until we have comparative figures for women of the same financial and social status, as college women differ from other women in many other respects than merely in the college training.

Effect of increasing n in the binomial expansion. It was noted in an earlier page that one effect of increasing n in the binomial expansion is to decrease the probabilities of numbers of successes within any specified interval of the mode. It was also noted that the ends of the distribution as a whole get smaller. These results can be more satisfactorily analyzed by means of the expression for σ just found. For $n = 4, 9$, and 16, $\sigma = c, 3c/2, 2c$. This value of σ really sums up the facts of the spread of these distributions better than the statements previously made, for any one who is familiar with the use of σ.

Digitized by Google

This method of approach also enables us to begin the discussion of what happens when n is made greater than any assigned number. If n is allowed to increase, c being kept constant, σ also increases, spreading the distribution thinner and thinner. A more satisfactory way of having the increase of n occur is to keep σ constant as n increases, by decreasing c. We have then $c\sqrt{npq} =$ a constant k, say, or c varies inversely as the square root of n. In the frequency polygon, the effect of this scheme is to crowd the ordinates closer together so that there is no longer a unit between them. Unless some compensating change is made, the result of such crowding will be to diminish the area of the polygon, and as the sum of the ordinates as they stand is in each case unity, the area when c is taken $= 1$ is also unity, and when c is less than one is less than unity. Now the area of the polygons is the same as that of the histogram, and is therefore $c \times$ [sum of ordinates]. If now we introduce a factor $1/c$ in each of the ordinates, the area will be the sum of the ordinates in their original form, which is 1. Denoting the height of the ordinate at any distance x from the mean by y, we have then

$$y = \frac{1}{c} \times \frac{n(n-1) \ldots (n-r+1)}{1 \times 2 \times 3 \times \ldots r} p^r q^{n-r}$$

$$= \frac{n(n-1) \ldots (nq+1-x/c)}{c \times 1 \times 2 \times 3 \times \ldots (np+x/c)} p^{np+x/c} q^{nq-x/c}$$

We have in the second expression replaced r by its equivalent $np + x/c$, and $n - r$ by $n - np - x/c = n(1-p) - x/c = nq - x/c$. This equivalence is true, because the distance from the "no success" ordinate to the general ordinate is either $r \times c$ or npc (distance to mean) $+ x$; these two expressions are therefore equal; $rc = npc + x$, and $r = np + x/c$.

If we find the ratio of this value of y to its value when $x = 0$, which we call y_o, then take logarithms, simplify somewhat, and then take limits as n becomes infinite, subject to the condition that $c^2npq = \sigma^2$, a constant, we find that

$$\text{Lim} \log y = \text{Lim} \log y_o - x^2/2\sigma^2$$

Digitized by Google

Now it can be shown also that

$$\text{Lim } y_o = \frac{1}{\sigma\sqrt{2\pi}}$$

Therefore we have

$$y = \frac{1}{\sqrt{2\pi} \times \sigma} e^{-x^2/2\sigma^2}$$

The curve with this equation is known in statistics as the *normal probability curve.* It is known also in the theory of error, with another constant M used instead of σ^2, the relation between them being $M^2 = 2\sigma^2$.

The derivation of this value of y from the binomial expansion was first carried through by A. de Moivre in the second edition of his *Doctrine of Chances* (1738), acknowledgment being there made of assistance received from Stirling in finding Lim log y_o, as given above. As Gauss and LaPlace each made great use of this equation in his work on theory of errors and probability, the law is often spoken of as the *Gauss-LaPlace law of frequency of error.* Unfortunately Gauss managed to give the impression that all departure from mean values must be according to this law, and thereby stifled further investigation for many years. As a matter of fact, this law results from much more general probability considerations than shown here, and by considering also the possibility that a given frequency distribution may be compounded of a large number of normal distributions with different means and σ's, we can get theoretical distributions which correspond closely to many actual distributions.

But there are many other actual distributions which can only awkwardly, or not at all, be so analyzed; a whole department of investigation has therefore sprung up based on the use of more general probability considerations or arbitrary substitutions of general parameters for the coefficients which actually occur at some point in the proofs or discussions connected with the normal curve. From the elementary point of view, to which we adhere in this book, the important thing is to realize that the normal curve is the simplest among many curves — simplest both in the underlying

Digitized by Google

logical foundation, and also in algebraic manipulation. Many tables of the values of the ordinates and areas of this curve are available, using either σ, M, or Q as the unit in the x-direction, and much use has been made of it in the construction of educational tests.

It will perhaps make the process of going to a limit seem more real to those not accustomed to algebraic thought, if we follow through in detail what happens to ordinates at this or that distance from the mean when we take n successively 4, 9, 16, and finally, infinite, that is to say, what it is for the normal curve. We can use the table on page 243, with certain modifications. In the first place, the mean occurs for number of successes, (r) 2, 4.5, and 8, respectively, and these positions are made the centers of reference. In the second place, we noted above that the values of y for the same values of x, rather than r, were comparable. If we take $\sigma = 1$, we have $c = 1$ when $n = 4$, $= 2/3$ when $n = 9$, and $= 1/2$ when $n = 16$, and as above, $x = (r - np)c$. In the third place, the ordinates must each be multiplied by $1/c$, to give the comparable values. With these modifications, we find the following corresponding ordinates:

s	-3	-1	0	1	3
$n = 4$	0.	.25	.375	.25	0.
$n = 9$	.003	.246	.....	.246	.003
$n = 16$	.0036	.244	.3924	.244	.0036
Normal curve	.0044	.242	.3989	.242	.0044
"Gamma" curve, $n = 4$	0.	.361	.391	.178	.015

The skew curve set down last does not fit in with the other curve or the polygons at all, it will be noted, except at the mean. But the process of gradual approach is obvious in the other figures.

Justification of the theory of least squares. The probability that a deviation will fall within a certain interval is the ratio of the area of the frequency curve between the boundaries of the interval, to the entire area of the frequency curve. The probability of an error x_1, i.e., of an error within an interval Δx_1, called the *least count* of the instrument, is approximately the area of a rectangle with base Δx_1 and height the ordinate of the frequency curve at

Digitized by Google

some point in the interval. If we assume that this frequency curve is due to chance variations from the mean, and may be represented correctly by the limit of a binomial expansion, the ordinate at x_1 is $\frac{1}{\sigma\sqrt{2\pi}} \times e^{-x_1^2/2\sigma^2}$. We have similar expressions for the ordinates at points x_2, x_3, etc. Then the probability that a set of errors x_1, x_2, x_3, etc., shall occur is the product of the probabilities that each separately occurs, that is to say, is

$$P = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-x_1^2/2\sigma^2} \cdot \Delta x_1 \times e^{-x_2^2/2\sigma^2} \cdot \Delta x_2 \ldots \times e^{-x_n^2/2\sigma^2} \cdot \Delta x_n$$

$$= \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-(x_1^2 + x_2^2 + \ldots + x_n^2)/2\sigma^2} \Delta x_1 \cdot \Delta x_2 \ldots \Delta x_n$$

Now it has been accepted as a principle of probability, that of alternative causes, that one is most probable on the basis of which the probability of the actual event is largest. If therefore we think of various sets of x's, each set depending on the choice of different means from which to measure the errors, or some other difference in standard with which the actual measurements are compared, that standard or mean is most probable for which the value of P is largest. The only change in P that can occur is in the X's in the numerator of the exponent of e. Now e raised to a negative power will be greatest when that exponent is least. The most probable set of causes then is that for which $x_1^2 + x_2^2 + \ldots + x_n^2$ is least. This statement is the basis of what is known as the method of least squares, which we have actually used in the determination of regression coefficients in the preceding chapter. If we have the simple question, given a set of x's which are direct measurements of a quantity, what is the best value of the quantity, on this theory? then we are merely asking, about what point as origin is the sum of the squares of the deviations least? It is left as a simple exercise for the student to show that that point is the arithmetic mean. And further, the sum of the squares of the deviations about the mean, which is less than a similar sum about any other point, is the basis for the standard deviation. We see then that the theory of least squares lends additional

Digitized by Google

prestige to two previous acquaintances, the arithmetic mean and the standard deviation.

Summary of Chapter XII. If n balls are drawn from an urn with constant probability p that any one ball is white, the probability that exactly r of the n balls will be white is given by the term in which the exponent of p is r, in the expansion of $(p + q)^n$.

The largest term in this expansion is that for which r is the integer between $np - q$ and $np + p$. If these expressions are both integers, the two corresponding terms are equal and larger than any other term.

On comparing probabilities for various values of n we note that:

(1) The probability of the most probable number of successes decreases as n is increased.

(2) The probabilities of numbers of successes within a fixed interval from the mode decrease as n is increased.

(3) The probability of n successes or n failures decreases as n is increased.

If we take the successive terms of the binomial expansion as ordinates of a frequency polygon, starting with the term corresponding to no successes and taking c as the distance between ordinates, we find that the distance from the first term to the mean is cnp, and $\sigma = c\sqrt{npq}$.

The formula $\sigma = c\sqrt{npq}$ makes it possible to estimate the probability that the proportion of white balls in the sample will reproduce, with any specified degree of accuracy, the proportion of white balls in the urn. It is therefore of fundamental importance, in the theory of probability, as it enables us to connect observed proportions with theoretical probability.

Mathematicians, scientists, and philosophers have verified the theory of probability by comparing the results of the theory with actual results in games of chance, biometry, theory of error, and actuarial theory.

The formula $\sigma = c\sqrt{npq}$ enables us to test whether certain differences in observed proportions are the result of chance alone.

Digitized by Google

If we increase n without limit, keeping the area under the frequency polygon constant and also keeping σ constant, the ordinates of the polygon approach as a limit the ordinates of the so-called normal probability curve, $y = \frac{1}{\sqrt{2\pi} \times \sigma} - e^{-x^2/2\sigma^2}$

Many actual frequency distributions can be represented approximately by a properly adjusted normal curve or by the sum of several such curves. Other frequency distributions can be represented much more satisfactorily by other curves. The development and study of such curves constitute part of the mathematical theory of statistics.

By the use of the normal curve, and the principle that the most probable of several causes is that for which the probability of the actual event is largest, it has been shown that the most probable "standard" is that for which the sum of the squares of the deviations about the standard is a minimum. We thus have an argument in favor of the use of the principle of least squares, introduced somewhat arbitrarily in earlier chapters, and see also more significance in the standard deviation.

EXERCISES

1. If three per cent of all children are left-handed, what is the probability that in a given group of 100 children none will be left-handed? What are the probabilities of one, two, three, four, five, six, seven, or eight left-handed children in such a group?

2. If analyzed past experience shows that 4 per cent of all insured white males of exact age 65 have died within a year, and it is found that 60 of a similar group of 1000 actually die within a year, should the group be regarded as essentially different from the general mass — that is to say, is the departure from the expected mortality greater than might be expected as a result of chance variation alone?

3. If 15 balls are drawn (with replacement after each draw) from an urn containing 30 white and 70 black balls, what is the most probable number of white balls? If 15 balls are drawn repeatedly, what is the average number of white balls per draw?

4. When a large corporation takes out for its employees a group insurance policy in a mutual insurance company, the excess premium above actual losses and expenses is, after the first year or two, returned as a dividend. The most probable dividend therefore depends on the average mortality of the group, and will be larger or smaller in a given year according as the actual mortality in any

Digitized by Google

year is above or below this average. As a large part of the premiums may be contributed by the employees, and the balance may be covered by the dividends, the financial risk assumed by the corporation in financing the plan depends on the stability of the mortality rate. It is necessary to consider the probability of specified variations in mortality such as the following:

If the average rate of mortality for a given group of about 10,000 has been about 10 per thousand, what is the probability that it will exceed 11 per thousand in a given year? (It is assumed that, because of the addition of new young people to the group and retirement of old people, the age distribution will not be substantially altered.)

SOLUTION: By formulas,

$$\text{expected (mean) mortality} = np = 100$$

$$\text{and} \qquad \sigma = \sqrt{npq} = \sqrt{99} = 9.95$$

A mortality rate of 11 per thousand implies 110 deaths, 10 or 1.005σ greater than the mean. By the table of areas of the normal curve, we see that this value of x corresponds to an area .842, and that the area further removed from the mean is .158. There is therefore about one chance in six that the rate of mortality will exceed 11 per thousand in a single year.

The student should determine the corresponding probabilities if the number of persons in the group is 3,000 or 100,000, and also if the variant probability is 10.5 or 12 per thousand rather than 11 per thousand.

Digitized by Google

CHAPTER XIII

PROBABLE ERROR

Meaning of probable error. The most important application of the theory of probability, from the point of view of the ordinary statistical worker, is connected with the idea of probable error. With many statistical constants, it is customary and usually desirable to state another, usually smaller, which is given the name "probable error" (P.E.). The real meaning of this term is nearly what we should expect from the previous discussion. It is that an error of the amount indicated, or larger, is as likely as not to have occurred. It is the quartile deviation of the frequency distribution for the constant which is being determined, the distribution which would be produced if a sample of the size of that actually taken were taken a very large number of times from material exactly like that from which the given sample was drawn. There is no significance in the notion of probable error unless there is a notion of sampling back of it. There is no "P.E." of a census result, where the entire population is enumerated rather than sampled. But in most of the results of statistical analysis, some idea of selection of a sample is inherent. An index number, for instance, is based on *some* rather than on *all* the possible series on which it might be based. A correlation table shows how closely two characteristics are associated in a limited number of individuals, and we run at once into a problem of sampling if we want to know how nearly the correlation found for this table holds in general. The value of the P.E. is usually written after the constant, in the form $\pm$ P.E. If we say, for instance, that the standard deviation of a certain series is 30 ± 2, we mean that the standard deviation of the sample is 30, and that if a large number of samples of the same size were selected, the quartile deviation of the frequency distribution of the σ's would be 2. It is not certain, of course, that the mean of this frequency distribution would be 30, but on the available information 30 is the best estimate we can make of

Digitized by Google

that mean. We accordingly say that the chances are even that the true standard deviation lies between 28 and 32.

Derivation of formula for the probable error of the mean. In order to help the student to understand the fundamental idea of the probable error, let us follow through the proof of the formula for the probable error of the mean.

Denote by n the number of items in each sample, and by κ the number of samples. Let us denote the frequencies of the values of the variable $x_1, x_2 \ldots x_r$ by $y_1, y_2 \ldots y_r$ in the first sample, by $y_1 + \delta y_1, y_2 + \delta y_2 \ldots y_r + \delta y_r$, in the second sample, and by similar expressions (with other values of δy_1, δy_2, etc.) in the third, fourth, and up to the κth sample. In this notation the mean of the first sample is given by:

$$nM = y_1 x_1 + y_2 x_2 + \ldots + y_r x_r$$

The mean for the second sample differs from this, and is given by:

$$n(M + \delta M_1) = (y_1 + \delta y_1)x_1 + (y_2 + \delta y_2)x_2 + \ldots + (y_r + \delta y_r)x_r$$

The means for the third, fourth, and other samples can be expressed in the same way, the only difference being that δy_1, δy_2, etc., take different values. Subtracting the expression for M found for the first sample from the several expressions for $M + \delta M$, we obtain the typical expression for δM:

$$\delta M = (x_1 \delta y_1 + x_2 \delta y_2 + \ldots + x_r \delta y_r)/n$$

Now the standard deviation of the frequency distribution of the various values of the mean is the root-mean-square of the deviations of these "means" about their mean. For convenience of notation, let us assume that that mean of the means is what we have called M; then the deviations are the various values of δM, and

$$\sigma_M = \sqrt{\frac{\Sigma(\delta M)^2}{\kappa}}$$

We have, then:

Digitized by Google

$$\kappa\sigma_M^2 = \frac{1}{n^2}\Sigma[x_1\delta y_1 + x_2\delta y_2 + \ldots + x_r\delta y_r]^2$$

$$= \frac{1}{n^2}\Sigma[x_1^2(\delta y_1)^2 + x_2^2(\delta y_2)^2 + \ldots + x_r^2(\delta y_r)^2 + 2x_1x_2\delta y_1\delta y_2 + 2x_1x_3\delta y_1\delta y_3 + \ldots]$$

As the summation applies to the δy's, not to the x's, the above may be simplified to:

$$\kappa\sigma_M^2 = \frac{1}{n^2}[x_1^2\Sigma(\delta y_1)^2 + x_2^2\Sigma(\delta y_2)^2 + \ldots + x_r^2\Sigma(\delta y_r)^2 + 2x_1x_2\Sigma\delta y_1\delta y_2 + 2x_1x_3\Sigma\delta y_1\delta y_3 + \ldots]$$

We shall next simplify further by deriving other values for these summations. If we assume that y_1 may be taken as the mean of all frequencies of the type $y_1 + \delta y_1$, then the various values of δy_1 are the deviations about this mean, and $\Sigma(\delta y_1)^2 = \kappa\sigma_{y_1}^2$, if we denote by σ_{y_1} the standard deviation of the frequencies $y_1 + \delta y_1$. But the probability that an item shall fall in the first group is the ratio of the average number of items in that group to the total number of items — in this case this ratio is y_1/n. By the formula $\sigma^2 = npq$, then, we have

$$\sigma_{y_1}^2 = n \times \frac{y_1}{n}\left(1 - \frac{y_1}{n}\right) = y_1\left(1 - \frac{y_1}{n}\right)$$

Therefore

$$\Sigma(\delta y_1)^2 = \kappa y_1\left(1 - \frac{y_1}{n}\right)$$

Again, when y_1 is changed to $y_1 + \delta y_1$, there result compensating changes in all the other y's. It is most natural to assume that the decrease in y_2, due to the increase in y_1, is the same proportion of the total change in y_1 as y_2 is of the total number of items in the sample outside of y_1. That is to say,

$$\text{part of } \delta y_2 = -\frac{y_2}{n - y_1} \times \delta y_1$$

Digitized by Google

Then $$\delta y_1 \times \delta y_2 = \frac{-y_2}{n-y_1}(\delta y_1)^2$$

plus another part, the product of δy_1 and changes in δy_2 due to other changes than δy_1. Therefore

$$\Sigma \delta y_1 \delta y_2 = \frac{-y_2}{n-y_1}\Sigma(\delta y_1)^2 + \text{another part.}$$

Now this summation, which is taken over the κ samples, is the product-sum term used in computing the coefficient of correlation between changes in y_1 and in y_2, and so is not affected by terms which are due to independent variation of y_1 and y_2; the "another part" term is therefore zero. Using the value found above for $\Sigma(\delta y_1)^2$, we have

$$\Sigma \delta y_1 \delta y_2 = \frac{-y_2}{n-y_1}\kappa y_1(1-y_1/n) = -\kappa y_1 y_2/n$$

Substituting these values in the expression for $\kappa\sigma_M^2$ above, we have:

$$n^2\kappa\sigma_M^2 = x_1^2\kappa y_1(1-y_1/n) + x_2^2\kappa y_2(1-y_2/n) + \ldots + x_r^2\kappa y_r(1-y_r/n) + 2x_1x_2(-\kappa y_1y_2/n) - 2\kappa x_1x_2y_1y_2/n \ldots$$

Multiplying out and rearranging terms, we have:

$$n^2\sigma_M^2 = x_1^2y_1 + x_2^2y_2 + \ldots + x_r^2y_r - \frac{1}{n}[x_1^2y_1^2 + x_2^2y_2^2 + 2x_1x_2y_1y_2 + \ldots + x_r^2y_r^2]$$

Now Σyx^2, the first part of the above right-hand side, is the summation to be used in finding the standard deviation of the original sample; we have exactly $\sigma^2 = \frac{\Sigma yx^2}{n} - M^2$, as the assumed mean is the origin. So $\Sigma yx^2 = n(\sigma^2 + M^2)$. The expression in brackets is clearly the square of nM. We have then:

$$n^2\sigma_M^2 = n(\sigma^2 + M^2) - \frac{1}{n}n^2M^2 = n\sigma^2, \text{ and therefore } \sigma_M = \sigma/\sqrt{n}$$

This is the standard deviation of the frequency distribution of the

Digitized by Google

means which would result on taking a large number of samples, and is sometimes called the standard error of the mean. If we assume that this frequency distribution will be normal, which seems a reasonable assumption, we may obtain the probable error or quartile deviation from it by multiplying by .6745. This is usually done, although the standard error is what is derived directly first, and if used as such no assumption is made as to normality.

Probable error of the difference of two means. An important problem is the determination of some criterion by which we may decide whether the observed difference between the means of two samples is really significant or is accidental, i.e., not greater than might arise by the fluctuations of random sampling from identical material. By the above proof, the standard error of the mean of the first sample, σ_M, is equal to $\sigma/\sqrt{n}$, where n is the number of cases in the first sample, and similarly for the second sample, $\sigma_{M'} = \sigma'/\sqrt{n'}$. Now if we took κ pairs of samples, with n items in the first sample of each pair, and n' in the second, and called the means of the first sample of each pair x, and of the second y, then by a formula proved at the end of Chapter X, the standard deviation of the difference of these means is given by $\sigma^2_{x-y} = \sigma^2_x + \sigma^2_y - 2r\sigma_x\sigma_y$. Now there is no correlation between the means of the paired samples, as they are taken independently; so $r = 0$. Substituting the value of σ_M for σ_x and $\sigma_{M'}$ for σ_y, as these are really the same, we have $\sigma^2_{x-y} = \sigma^2_M + \sigma^2_{M'} = \sigma^2/n + \sigma'^2/n'$. The standard error of the difference of two means is therefore $\sqrt{[\sigma^2/n + \sigma'^2/n']}$, and the probable error .6745 times this.

This formula is of great value. Consider the following illustration. Suppose that 1000 men in New York are selected at random and measured, and have an average height of 5 feet, 6 inches, with a σ of 4 inches, and that 2000 men in Chicago, selected and measured in the same way, have an average height of 5 feet, 7 inches, with a σ of 4.2 inches. Is it fair to say that the men in Chicago are taller than those of New York, or would the selection of larger samples change the result? The formula just proved says that the standard error of the difference in height under these cir-

Digitized by Google

cumstances is $\sqrt{[16/1000 + 17.64/2000]} = .158$ inches. The actual difference is therefore over 6 times the standard error; as long as it is over 3 times, chance occurrence is less than once in 400 trials; under the given conditions we can therefore feel certain that there is a real difference between the height of men in the two cities. The amount of the difference, using the P.E. with the notation explained above, is $1 \pm .1$ inch. If the samples had been 100 and 200, respectively, instead of 1000 and 2000, the standard error would have been .498, and the actual difference would then have been only twice the standard error, and therefore of doubtful significance.

Formulas for standard errors of statistical constants. The following list gives the standard errors of the more important statistical constants. The probable error is obtained in each case by multiplying by .6745, on the assumption of normal distribution. For most purposes it is accurate enough to multiply by 2/3; except for the fact that the use of the probable error is usual, the standard errors serve every purpose.

Binomial distribution	$\sqrt{npq}$
Any compartment of a frequency distribution	$\sqrt{y_1(1 - y_1/n)}$
The arithmetic mean	$\sigma/\sqrt{n}$
The median	$5\sigma/4\sqrt{n}$
The difference of two means	$\sqrt{[\sigma^2/n + \sigma'^2/n']}$
The standard deviation	$\sigma/\sqrt{2n}$ for normal distribution; $\sqrt{\left[\frac{m_4 - \sigma^4}{4n\sigma^2} \cdots\right]}$ for any distribution, where m_4 is the fourth moment, the sum of the fourth powers of the deviations divided by n.
The coefficient of correlation r	$(1 - r^2)/\sqrt{n}$
Coefficient of variation v	$\frac{v}{\sqrt{2n}}\{1 + 2(v/100)^2\}^{1/2}$

Digitized by Google

Coefficient of regression m_x $\quad \dfrac{\sigma_x\sqrt{(1-r^2)}}{\sigma_y\sqrt{n}}$

Correlation ratio η , $\quad (1-\eta^2)/\sqrt{n}$

$\eta^2 - r^2$, the test for the linearity of regression

$$2\sqrt{\frac{\eta^2-r^2}{n}}\sqrt{[1+(1-\eta^2)^2-(1-r^2)^2]}$$

A probability basis for the coefficient of correlation. Professor H. L. Rietz has recently shown[1] that in a certain simple problem in probability, a simple and informing interpretation may be given to the coefficient of correlation. Suppose that ten dice are thrown, and the number of sixes noted. Call the number of sixes on the first throw α_1, on the second α_2, and so on. Before the second throw is made, select at random four of the dice thrown the first time and throw them again. Denote by β_1 the total number of sixes among the ten dice — six as thrown originally and four as rethrown. After the second throw, again rethrow four dice, denoting by β_2 the number of sixes. We thus derive two series of numbers, in which there is a definite pairing of any number in one series with a number in the other series. If now we compute the coefficient of correlation for these two series α and β, we find that it is .6, that is to say, the proportion of common elements in the method of throwing by which the two series are derived. In general it may be shown that if s dice remain unchanged and r dice are rethrown, the coefficient of correlation is $s/(r+s)$.

Attention is called to the interpretation of the coefficient of correlation in the preceding problem as "the proportion of common elements." This idea is a good one to have in mind when thinking of the significance of some particular application of correlation. It must not be thought to apply too generally, however, as in actual problems the conditions are not so simple. The probabilities on successive throws may be unequal, or the selection of dice to be rethrown may depend in part on whether or not they are sixes; or the conditions may in other ways depart from regularity. In

[1] *Annals of Mathematics*, vol. 21, 1920, "Urn Schemata as a Basis for the Development of Correlation Theory."

Digitized by Google

such cases the coefficient of correlation will not correspond exactly to the simple proportion stated.

Summary of Chapter XIII. The probable error of a statistical result is the variation which is as likely as not to occur if the same result is determined for another sample of the same size as the first sample, constituted of items like those in the first sample.

On certain reasonable assumptions, it is shown that the probable error of the arithmetic mean is equal to the standard deviation of the sample, divided by the square root of the number of cases in the sample.

The probable error of the difference of the means determined from two samples drawn from identical material is equal to $\sqrt{[\sigma^2/n + \sigma'^2/n']}$, where σ, σ' are the standard deviations of the two samples and n, n' the number of items in the two samples. Unless the actual difference of the means exceeds three times this probable error, it should not be considered significant.

Certain approximate formulas for probable errors are generally accepted.

In certain simple cases, it may be shown that the coefficient of correlation is the proportion of common elements in the causes of two series of events.

The "probable error," as explained above, is the error that is to be expected according to the theory of probability, on the basis of the number of cases and the magnitude of the constants. It has nothing to do with the errors due to inadequate or defective data, incorrect computation, or slip-shod thinking in regard to the problem which is being analyzed.

EXERCISE

The results of a set of Brown University Psychological Tests given to certain High School seniors were as follows:

	Number Tested	Mean Scores	Standard Deviations
Boys	1262	48.9	11.9
Girls	2071	43.6	11.6

On the basis of these figures, is the difference between the boys and girls significant?

Digitized by Google

CHAPTER XIV

THREE AND FOUR VARIABLE CORRELATION

We have seen, in Chapters VIII and X, that in certain cases the relation between two characteristics of a mass of items may be summarized in a definite algebraic formula. It was found, for instance, that for sixty years in Ohio,

Yield of corn in bushels per acre = 27 + 1.7 times July rainfall in inches.

It is obvious on reflection that such a formula assumes that other conditions than July rainfall have about their average value, and that a more complete formula would take account of variations in such other conditions as would most alter the results, as for instance amount of sunshine, fertilization, rainfall in other months, expected price per bushel, and availability of labor. The purpose of this chapter is to develop methods for analyzing relationships between three or four characteristics of a mass of items. We shall find, among other things, a regression equation for the most probable value of one characteristic when the values of two others are given, and a measure of relationship, called the multiple coefficient of correlation, which is really the coefficient of correlation between these most probable values and the actual values.

A problem in three variable relationship; assembling the data. In order to illustrate the analysis of a three variable problem, let us try to discover how good a prediction of success in the last three years of college work can be based on the record of work in the first year and a psychological test score. In order to throw light on this problem, the following two-way frequency tables were compiled at the suggestion of the writer, from the records of the 174 members of the class of 1923 at Brown University who had a psychological test score, a freshman year academic average, and an academic average after the freshman year. Students who failed to return after the freshman year were thus excluded; as

Digitized by Google

many such students had low test scores and also low freshman averages, the correlation between these two quantities is less than if all freshmen had been included.

TABLE 1. BASED ON ENTIRE GROUP

Test Scores	Academic Average after Freshman Year									Total
	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85–89	90–94	
80–89.....							1	2	1	4
70–79.....		1			8	9	12	9	3	42
60–69.....	1	2	4	1	4	24	12	12	4	64
50–59.....	1		3	2	8	13	13	3		43
40–49.....	1	1	1	1	4	5	3	1	2	19
30–39.....		1	1							2
Total.....	3	5	9	4	24	51	41	27	10	174

TABLE 2. BASED ON ENTIRE GROUP

Test Scores	Academic Average Freshman Year							Total
	55–59	60–64	65–69	70–74	75–79	80–84	85–89	
80–89.....			1			2	1	4
70–79.....		1	9	3	12	11	6	42
60–69.....	2	2	15	10	15	11	9	64
50–59.....		10	8	10	8	6	1	43
40–49.....		6	3	4	2	1	3	19
30–39.....		1	1					2
Total.....	2	20	37	27	37	31	20	174

TABLE 3. BASED ON ENTIRE GROUP

Academic Av. Freshman Year	Academic Average after Freshman Year									Total
	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85–89	90–94	
85–89.....					1		2	9	8	20
80–84.....						5	16	8	2	31
75–79.....			2	1	2	15	10	7		37
70–74.....	1	1	1	1	3	8	9	3		27
65–69.....	1	1	3	1	11	17	3			37
60–64.....	1	2	2	1	7	6	1			20
55–59.....		1	1							2
Total.....	3	5	9	4	24	51	41	27	10	174

Digitized by Google

TABLE 4. BASED ON THOSE WITH TEST SCORES OF 70 OR OVER

ACADEMIC AV. FRESHMAN YEAR	ACADEMIC AVERAGE AFTER FRESHMAN YEAR								TOTAL
	55–59	60–64	65–69	70–74	75–79	80–84	85–89	90–94	
85–89.....				1			2	4	7
80–84.....					2	7	4		13
75–79.....				1	4	2	5		12
70–74.....					1	2			3
65–69.....				6	2	2			10
60–64.....	1								1
Total...	1			8	9	13	11	4	46

TABLE 5. BASED ON THOSE WITH TEST SCORES OF 60–69.9

ACADEMIC AV. FRESHMAN YEAR	ACADEMIC AVERAGE AFTER FRESHMAN YEAR								TOTAL
	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85–89	90–94	
85–89.....							1	6	2	9
80–84.....						1	6	2	2	11
75–79.....			1		1	9	2	2		15
70–74.....		1		1		3	3	2		10
65–69.....	1		2		2	10				15
60–64.....					1	1				2
55–59.....		1	1							2
Total.....	1	2	4	1	4	24	12	12	4	64

TABLE 6. BASED ON THOSE WITH TEST SCORES OF UNDER 60

ACADEMIC AV. FRESHMAN YEAR	ACADEMIC AVERAGE AFTER FRESHMAN YEAR								TOTAL
	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85–89	90–94	
85–89.....							1	1	2	4
80–84.....						2	3	2		7
75–79.....			1	1		2	6			10
70–74.....	1		1		3	4	4	1		14
65–69.....		1	1	1	3	5	1			12
60–64.....	1	1	2	1	6	5	2			17
Total.....	2	2	5	3	12	18	16	4	2	64

The last three tables were found in order to make possible a further investigation of the close relation obviously indicated by Table 3 between averages after and during the freshman year. The question naturally arises whether the cases in which these

Digitized by Google

averages are markedly different are those with unusual test scores. Tables 4, 5, and 6 make possible a discussion of this question.

In view of previous discussions, one method which suggests itself for summarizing these facts is to find the coefficients of correlation, standard deviations, and regression equations. As the student may verify, they are as follows:

Coefficients of correlation:

Between test scores and post-freshman averages (r_{13}) =	.275
Between test scores and freshman averages (r_{23}) =	.317
Between freshman and post-freshman averages (r_{12}) =	.711
Preceding for those with test scores of 70 or over	.679
Preceding for those with test scores of 60–69.9	.671
Preceding for those with test scores of under 60	.545

Standard deviations:

	All Cases	High Test	Av. Test	Low Test
Post-freshman averages (x_1)	8.21	7.01	8.67	8.70
Freshman averages (x_2)	7.59	7.11	7.93	7.72
Test scores (x_3)	5.18			

Regression equations:

(All test scores)	$x_1 = 78 + .77\,(x_2 - 75)$	σ of estimate = 5.8
(High test scores)	$x_1 = 79 + .67\,(x_2 - 75)$	σ of estimate = 5.2
(Av. test scores)	$x_1 = 79 + .73\,(x_2 - 75)$	σ of estimate = 6.4
(Low test scores)	$x_1 = 78 + .62\,(x_2 - 75)$	σ of estimate = 7.3
	$x_1 = 77 + .44\,(x_3 - 60)$	σ of estimate = 7.9
	$x_2 = 73 + .58\,(x_3 - 60)$	σ of estimate = 7.2

These regression equations predict a later event in terms of an earlier one; the other equations work backward chronologically and therefore are not of sufficient interest to state.

Discussion of these equations and correlations. It is evident that the test score throws little additional light on the probable post-freshman average, after the freshman average is known. It is true that on the whole the men with high test scores are likely to do better than others, both as freshmen and later; but there are so many exceptions and the degree of differentiation is relatively so

Digitized by Google

small, that it would not be desirable to lay much emphasis on this fact.

The smallness of the differences between the four first-stated regression equations, in spite of the apparent significance of test scores implied by the last two equations, indicates that perhaps the method does not bring out all the relations really involved in the tables. At any rate, the method is not as convenient and simple as it might be, in so far as it really implies a sort of three-way tabulation. We shall derive a few pages further along a method which is somewhat more direct and simple, when it applies.

Application of percentage method of analysis. Before taking up this more careful method, however, it may be worth while to summarize these two-way tables in a way that might well occur to a man with the analytic type of mind who had never heard of correlation. It is clear that in general a student's academic average after freshman year falls in the same class as or next higher class than his average in freshman year. A measure of the closeness of this relationship may be obtained by noting the percentage of students in each group whose two grades differ by more than this amount. These percentages are as follows:

All test scores	46
High test scores	35
Av. test scores	47
Low test scores	55

That is to say, the correspondence is closest in the case of those with high scores and lowest in the case of those with low scores.

Regression equation in three variables. Let us now proceed to derive the necessary formulas for the more simple and direct method mentioned above.

In Chapter VIII it was shown how to find the best value of m so that $y = mx$ shall be the equation of the line of closest fit to the points corresponding to the given pairs of values of x and y. "Closest fit" was defined by the condition that the sum $\Sigma(y - mx)^2$ shall be a minimum; this definition was justified in Chapter XII

Digitized by Google

by showing that in the case of normal distribution the most probable cause was that for which the sum of the squares of the deviations was a minimum.

Corresponding to $y = mx$ or $x = m'y$ as the "best" formulas of the first degree connecting two variables x and y, are similar "best" formulas of the first degree connecting three, four, or more variables. In deriving these equations and simplifying the methods of computation, we shall find it desirable to use substantially the notation suggested by Yule. Consider first the case of one variable as a function of two others. The three variables are called x_1, x_2, and x_3, and the coefficients m_{12-3}, m_{21-3}, m_{13-2}, etc., where the first subscript is the same as that of the variable which is being expressed in terms of the others, the second is that of the variable of which this m is the coefficient, and the subscripts following the $-$ are those of the other variables which are being considered. When it is perfectly clear what is meant, we shall for simplicity drop the subscripts after the $-$, but shall retain them where confusion might result. Thus m_{12-3} will be written m_{12} where there is no danger of confusing it with the real m_{12}, i.e., the coefficient in the two variable case, $x_1 = m_{12}x_2$, or with m_{12-34} or other m's of that sort. The regression equations in three variables are then:

$$x_1 = m_{12-3}x_2 + m_{13-2}x_3 + c_1$$
$$x_2 = m_{21-3}x_1 + m_{23-1}x_3 + c_2$$
$$x_3 = m_{31-2}x_1 + m_{32-1}x_2 + c_3$$

Let us now, as in the two-variable case, apply the condition that the algebraic sum of the errors shall be zero. That is to say, we want

$$\Sigma(x_1 - m_{12}x_2 - m_{13}x_3 - c_1) = 0, \text{ or } \Sigma x_1 - m_{12}\Sigma x_2 - m_{13}\Sigma x_3 - Nc_1 = 0$$

Now $\Sigma x_1 = N\bar{x}_1$, $\Sigma x_2 = N\bar{x}_2$ and $\Sigma x_3 = N\bar{x}_3$

so substituting, we have

$$c_1 = \bar{x}_1 - m_{12}\bar{x}_2 - m_{13}\bar{x}_3$$

and so our original equation becomes:

Digitized by Google

$$x_1 = m_{12}x_2 + m_{13}x_3 + \bar{x}_1 - m_{12}\bar{x}_2 - m_{13}\bar{x}_3$$

or $$x_1 - \bar{x}_1 = m_{12}(x_2 - \bar{x}_2) + m_{13}(x_3 - \bar{x}_3)$$

Using capital letters to denote deviations from the mean, the equation becomes: $X_1 = m_{12}X_2 + m_{13}X_3$ and similarly for the other two equations.

As in the two-variable case, we now use the condition that the sum of the squares of the deviations shall be a minimum, i.e., $\Sigma(X_1 - m_{12}X_2 - m_{13}X_3)^2$ shall be a minimum. This minimum is denoted by the symbol $N\sigma^2_{1-23}$, thus defining standard deviations of a higher order.

Expanding, we have

$$\Sigma(X_1^2 + m_{12}^2X_2^2 + m_{13}^2X_3^2 - 2m_{12}X_1X_2 - 2m_{13}X_1X_3 + 2m_{12}m_{13}X_2X_3)$$

Now by formulas derived in Chapters VII and X, with obvious change in notation:

$$\Sigma X_i^2 = N\sigma_i^2 \quad \text{and} \quad \Sigma X_iX_j = Nr_{ij}\sigma_i\sigma_j$$

Substituting these formulas and dropping the factor N, we have:

$$\sigma'^2_{1-23} = \sigma_1^2 + m_{12}^2\sigma_2^2 + m_{13}^2\sigma_3^2 - 2m_{12}r_{12}\sigma_1\sigma_2 - 2m_{13}r_{13}\sigma_1\sigma_3 + 2m_{12}m_{13}r_{23}\sigma_2\sigma_3$$

Now as previously noted in Chapter VIII, page 163, the expression

$$lx^2 + 2txy + my^2 + 2ux + 2vy + s$$

is proved in calculus to have a minimum when

$$lx + ty + u = 0 \quad \text{and} \quad tx + my + v = 0$$

provided that l and m are both positive and lm is greater than t^2. In this case, then, the values of m_{12} and m_{13} for which σ'^2_{1-23} has a minimum are given by the equations:

$$\sigma_2^2m_{12} + r_{23}\sigma_2\sigma_3m_{13} - r_{12}\sigma_1\sigma_2 = 0$$

$$r_{23}\sigma_2\sigma_3m_{12} + \sigma_3^2m_{13} - r_{13}\sigma_1\sigma_3 = 0$$

provided $\sigma_2^2\sigma_3^2$ is greater than $(r_{23}\sigma_2\sigma_3)^2$, which is obviously true.

Digitized by Google

Dividing through by common factors and transposing, we may put these equations in the form:

$$\sigma_2 m_{12} + r_{23}\sigma_3 m_{13} = r_{12}\sigma_1$$
$$\sigma_2 r_{23} m_{12} + \sigma_3 m_{13} = r_{13}\sigma_1$$

Solving, we have:

$$m_{12} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \quad \frac{\sigma_1}{\sigma_2} \quad \text{and} \quad m_{13} = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \quad \frac{\sigma_1}{\sigma_3}$$

The corresponding regression equation using these coefficients is:

$$X_1/\sigma_1 = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \quad \frac{X_2}{\sigma_2} + \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \quad \frac{X_3}{\sigma_3}$$

By obvious shifting of subscripts, we may obtain two other regression equations expressing X_2 and X_3, respectively, in terms of the other variables.

Determination of value of σ_{1-23}^2. As we may write

$$lx^2 + 2txy + my^2 + 2ux + 2vy + s$$

in the form $\quad x(lx + ty + u) + y(tx + my + v) + ux + vy + s$

it is clear that the minimum value of the expression, which occurs when

$$lx + ty + u = 0 \quad \text{and} \quad tx + my + v = 0$$

is given by $ux + vy + s$. So in this case, the minimum value of σ'^2_{1-23}, that is to say, σ_{1-23}^2, is given by:

$$\sigma_{1-23}^2 = -r_{12}\sigma_1\sigma_2 m_{12} - r_{13}\sigma_1\sigma_3 m_{13} + \sigma_1^2$$

$$\sigma_{1-23}^2/\sigma_1^2 = -r_{12}\left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right) - r_{13}\left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right) + 1$$

$$= \frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

The expression σ_{1-23}^2 for the average of the squares of the

Digitized by Google

deviations of X_1 from the values computed by the formula $m_{12}X_2 + m_{13}X_3$ is a valuable measure of the accuracy of fit of the equation and corresponds to the expression $\sigma^2(1-r^2)$ in the two-variable case. In the next section we shall alter the form somewhat, in a way that will make the analogy closer.

Multiple coefficient of correlation. The multiple coefficient of correlation is defined to be the coefficient of correlation between the observed X_1 and the X_1' that would be predicted from the right-hand side of the preceding regression equation, and is denoted by the symbol R_{1-23}. We have then:

$$R_{1-23} = \Sigma X_1 X_1'/N\sigma_1\sigma_{X_1'}$$

Now clearly

$$\sigma^2_{X_1'} = \Sigma(m_{12}X_2 + m_{13}X_3)^2/N = m^2_{12}\sigma^2_2 + m^2_{13}\sigma^2_3 + 2m_{12}m_{13}r_{23}\sigma_3\sigma_2$$
$$= m_{12}(m_{12}\sigma^2_2 + m_{13}r_{23}\sigma_3\sigma_2) + m_{13}(m_{13}\sigma^2_3 + m_{12}r_{23}\sigma_2\sigma_3)$$

Using the two equations proved above, this becomes:

$$\sigma^2_{X_1'} = r_{12}m_{12}\sigma_1\sigma_2 + r_{13}m_{12}\sigma_1\sigma_3 = \sigma^2_1 - \sigma^2_{1-23}$$

$$\sigma^2_{X_1'} = \sigma^2_1\left(\frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1-r^2_{23}}\right)$$

We have also:

$$\Sigma X_1 X_1'/N = m_{12}\Sigma X_1X_2/N + m_{13}\Sigma X_1X_3/N$$
$$= m_{12}r_{12}\sigma_1\sigma_2 + m_{13}r_{13}\sigma_1\sigma_3 = \sigma^2_{X_1'}$$

We have therefore:

$$R_{1-23} = \sigma^2_{X_1'}/\sigma_1\sigma_{X_1'} = \sigma_{X_1'}/\sigma_1 = \sqrt{\frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1-r^2_{23}}}$$

or

$$= \sqrt{(1-\sigma^2_{1-23}/\sigma^2_1)}$$

This last equation may also be put in the form

$$\sigma_{1-23} = \sigma_1\sqrt{(1-R^2_{1-23})}$$

Digitized by Google

Partial correlation coefficients. In Chapter X, it was shown that the ordinary coefficient of correlation is the geometric mean of the coefficients of regression of x on y and of y on x. In analogy to this property, which could be used as a definition, the partial coefficient of correlation between two variables x_i and x_j, is defined to be the square root of the product of the coefficient of x_j in the regression equation for x_i in terms of other variables, multiplied by the coefficient of x_i in the regression equation for x_j. Or, in symbols:

$$r^2_{ij-k} = m_{ij-k} \times m_{ji-k}$$

Substituting values as found above, this may be simplified as follows:

$$r^2_{12-3} = \frac{\sigma_1(r_{12} - r_{13}r_{23})}{\sigma_2(1-r^2_{23})} \times \frac{\sigma_2(r_{12} - r_{13}r_{23})}{\sigma_1(1-r^2_{13})}$$

$$r_{12-3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r^2_{23})(1-r^2_{13})}}$$

In words, the partial coefficient of correlation may be said to measure the degree of relationship between x_1 and x_2, with the effect of x_3 considered but eliminated. If x_3 is altogether independent of x_1 and x_2, as measured by the ordinary coefficients of correlation, $r_{13} = 0$, $r_{23} = 0$, and $r_{12} = r_{12-3}$. In the more usual case, however, the difference between r_{12} and r_{12-3} is considerable, in fact, they may even be of opposite sign. In such cases, the partial coefficient of correlation is a better measure of the true degree of relationship, as it shows what the correlation would be in a group of cases for which the third variable is constant. The main application of partial correlation is thus testing what the correlation between two variables would be if the third variable were not interfering with the apparent relationships.

In the example considered at the beginning of this chapter, the coefficients of partial correlation are as follows, in comparison with the two-variable coefficients:

$$r_{12-3} = .684 \quad (r_{12} = .711)$$
$$r_{13-2} = .075 \quad (r_{13} = .275)$$
$$r_{23-1} = .180 \quad (r_{23} = .317)$$

Digitized by Google

That is to say, for the group discussed, there is very little real connection between test scores and academic grades, either during or after the freshman year; the apparent effect indicated by the two-variable correlations is incidental to the high correlation between grades during and after the freshman year.

Three-variable regression equation for preceding example. In the writer's experience, however, partial correlations are not as valuable as the coefficient of multiple correlation and the three-variable regression equation. In the example considered at the beginning of this chapter, the important regression equation, expressing grades *after* freshman year in terms of the other two variables, is:

$$x_1 = 79 + .75\ (x_2 - 75) + .09\ (x_3 - 65)$$

The corresponding coefficient of multiple correlation is .713, and the standard deviation of estimates based on this equation is 5.75.

Geometric interpretation of multiple and partial correlation. Just as the two-variable regression equation may be represented graphically by a straight line, so the three-variable regression equation corresponds to a plane. The intersection of the plane $x_1 = m_{12 \cdot 3}x_2 + m_{13 \cdot 2}x_3$ with the plane $x_3 = k$ is a straight line, and the slope of this straight line is the same for all values of k, namely, $m_{12 \cdot 3}$. We therefore see that $m_{12 \cdot 3}$ is the rate of increase of x_1 per unit increase of x_2 when x_3 is kept constant. The constancy of this rate as given by this equation is of course not necessarily a truly correct summary of the actual facts, but is a sort of average of what would be found if separate rates could be found for each value of x_3.

Each of the three regression equations corresponds to a plane passing through the origin, that is to say, the point corresponding to the mean value of each variable. In the case of perfect correlation between all three variables, these planes would all coincide; their departure from coincidence is one of the possible measures of the extent to which they are not correlated.

An interesting interpretation of partial and multiple correlation

Digitized by Google

is available to those who are familiar with spherical trigonometry. Perhaps the simplest approach is as follows:

As correlation coefficients vary in value from -1 to $+1$, any correlation coefficient can be stated to be the cosine of some angle. Let us then take the three first order coefficients in a three-variable problem to be the cosines of the sides of a spherical triangle. This assumes, to be sure, that certain limitations in values possible exist, but it can be shown that all such limitations necessarily hold for three such correlation coefficients. We have, let us say, then:

$$r_{12} = \cos a, \qquad r_{13} = \cos b, \qquad r_{23} = \cos c$$

As proved above,

$$r_{12 \cdot 3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

which becomes, substituting values,

$$\frac{\cos a - \cos b \cos c}{\sqrt{(1 - \cos^2 b)(1 - \cos^2 c)}} = \frac{\cos a - \cos b \cos c}{\sin b \sin c} = \cos A$$

That is to say, the partial coefficient of correlation is the cosine of the angle opposite the side corresponding to the first order coefficient.

Similarly, it was proved above that

$$R_{1 \cdot 23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2} = \frac{\cos^2 a + \cos^2 b - 2 \cos a \cos b \cos c}{1 - \cos^2 c}$$

$$= \frac{\cos^2 a + \cos^2 b - 2 \cos a \cos b \cos c}{\sin^2 c}$$

Now, if we denote by $p_{1 \cdot 23}$ the altitude of the spherical triangle perpendicular to side c, we have:

$$\sin p_{1 \cdot 23} = \sin a \sin B, \text{ and } \cos^2 p_{1 \cdot 23} = 1 - \sin^2 a \sin^2 B$$

$$\cos^2 p_{1 \cdot 23} = 1 - \sin^2 a (1 - \cos^2 B) = 1 - \sin^2 a + \sin^2 a \cos^2 B$$

$$= \cos^2 a + \sin^2 a \frac{(\cos b - \cos c \cos a)^2}{\sin^2 a \sin^2 c}$$

Digitized by Google

$$= \cos^2 a + \frac{\cos^2 b - 2\cos a \cos b \cos c + \cos^2 a \cos^2 c}{\sin^2 c}$$

$$= \frac{\cos^2 a \sin^2 c + \cos^2 a \cos^2 c + \cos^2 b - 2\cos a \cos b \cos c}{\sin^2 c}$$

$$= \frac{\cos^2 a + \cos^2 b - 2\cos a \cos b \cos c}{\sin^2 c}$$

Therefore $R^2_{1-23} = \cos^2 p_{1-23}$, or the coefficient of multiple correlation is the cosine of the altitude of the spherical triangle perpendicular to the side corresponding to the r which appears in the denominator of the formula for R.

From this relation, it is obvious that R_{1-23} can never be less than r_{12} or r_{13}. For in any spherical triangle with sides less than 90°, any altitude is equal to or less than either of the sides meeting at the vertex from which it is drawn. The cosine of the altitude is therefore equal to or greater than the cosine of either side meeting at that vertex. Hence the theorem follows.

This theorem also follows from the algebraic formula for R_{1-23}, which may be put in the form:

$$R^2_{1-23} = r^2_{12} + \frac{(r_{13} - r_{12}r_{23})^2}{1 - r^2_{23}}$$

From this form it is obvious that R_{1-23} is greater than r_{13}; and a similar form makes it obvious that it is greater than r_{12}.

Application of multiple correlation to historical series. An important application of multiple correlation is to the prediction of one economic variable in terms of two others. If we take x_2 and x_3 as of an earlier date than x_1, and have the good fortune to find that R_{1-23} is reasonably large, on the basis of past experience, it is obvious that the regression equation for x_1 permits predictions of future values of x_1 on the basis of current values of x_2 and x_3. Professor H. L. Moore, in his path-finding application of this method, for instance, found that a prediction of the yield of cotton in terms of May rainfall and June temperature was more accurate

Digitized by Google

than the prediction made by the Federal Department of Agriculture at the end of August, on the basis of thousands of reports from observers. The test of accuracy is the "spread" between actual and predicted — in the case of the correlation estimate, $\sigma_1\sqrt{(1 - R^2_{1-23})}$. The details of this study are given in Professor Moore's monograph, *Forecasting the Yield and Price of Cotton.* Professor Moore obtains also an equation giving change of price in terms of change in yield and in general price-level.

Professor Moore's pupils and others have applied this method to a number of agricultural products and have made further refinements in the analysis of cotton production and price.

An important caution in the study of time series by multiple correlation is, always determine the spread of estimate $[\sigma_1(1 - R^2)]$ before making any judgment in regard to the accuracy of estimate or the relative desirability of different variations in method. In particular, the coefficient of correlation will usually seem rather small if the correlation is found between differences or changes, and rather large if found between the quantities themselves. The proper way to decide between these two methods in a particular case is to ignore the obvious comparison between correlation coefficients and compare the spreads of the estimates resulting from the two methods.

Four-variable correlation. By exactly the same methods as used for three variables, it is possible to obtain equations expressing the relations existing between four or more variables. To determine the coefficients in the equation

$$x_1 = m_{12-34}x_2 + m_{13-24}x_3 + m_{14-23}x_4$$

we have the following simultaneous equations:

$$\begin{aligned}
\sigma_2 m_{12} + r_{23}\sigma_3 m_{13} + r_{24}\sigma_4 m_{14} &= r_{12}\sigma_1 \\
\sigma_2 r_{23} m_{12} + \sigma_3 m_{13} + r_{34}\sigma_4 m_{14} &= r_{13}\sigma_1 \\
\sigma_2 r_{24} m_{12} + r_{34}\sigma_3 m_{13} + \sigma_4 m_{14} &= r_{14}\sigma_1
\end{aligned}$$

The coefficient of correlation between the actual x_1 and the X_1 resulting from this equation is the multiple coefficient of correlation R_{1-234}, which is given by the formula: $R^2_{1-234} = N/D$, where

Digitized by Google

$$N = r_{12}^2 + r_{13}^2 + r_{14}^2 - 2r_{12}r_{13}r_{23} - 2r_{12}r_{14}r_{24} - 2r_{13}r_{14}r_{34}$$
$$+ 2r_{12}r_{13}r_{24}r_{34} + 2r_{12}r_{14}r_{23}r_{34} + 2r_{13}r_{14}r_{23}r_{24} - r_{12}^2r_{34}^2 - r_{13}^2r_{24}^2 - r_{14}^2r_{23}^2$$

and $$D = 1 - r_{23}^2 - r_{24}^2 - r_{34}^2 + 2r_{23}r_{24}r_{34}$$

As before, the spread of the actual as compared with the computed x_1's is given by σ_{1-234}, where $\sigma_{1-234}^2 = \sigma_1^2 (1 - R_{1-234}^2)$

Four-variable regression equations should, in the opinion of the writer, be used only with considerable caution, and the use of five or more variables is to be discouraged still more. At the best, these formulas give only an average relationship, the best there is, to be sure, of the first degree algebraically. But the assumption of the adequacy of a first-degree formula becomes increasingly dangerous as we increase the number of variables. It will often be found, moreover, that the increase in accuracy of the prediction based on the larger number of variables is negligible. The recommended procedure is to construct the equation in three variables, using for x_2 and x_3 those which correlate most highly with x_1. Then the variable which has the next highest correlation may be tried out as x_4, to the extent at least of seeing whether R_{1-234} is appreciably larger than R_{1-23}. If not, try other possibilities for x_4; if so, and other variables not yet tried seem to be worth consideration, try one of them as an x_4 before working it in as a fifth variable. Perhaps it gives no new information not already covered. Or further examination of the data may show that it is possible to break them up into several classes, in each of which one of the variables is substantially constant.

In general, the writer believes that an intensive examination of two- or three-variable correlations, with full consideration of possible non-linear relationships, will be found more valuable than the cursory inspection implied by regression equations in four or more variables. A method for carrying out a more intensive examination of the three-variable case is indicated in the next section.

A non-linear three-variable regression equation. Suppose we approach the problem of determining three-variable relationships

Digitized by Google

by thinking what changes in the two-variable equation would naturally arise from the attempt to take account of the third variable. Suppose for the sake of definiteness we say that we are trying to express men's weight x_1 as a function of height x_2 first, and then of height and age x_3 combined. We might obtain a general formula for all the men considered, no matter what the age, giving say, $x_1 = mx_2 + b$. Or we might obtain similar formulas for the men of each age group. On the face of it, the values of m and b would probably be different for the different groups. The constant m, for instance, gives the increase in weight per unit increase in height. This might well be somewhat larger for the older men than for the younger. The constant b, gives the weight of a man of zero height, as predicted from the heights and weights actually observed. More naturally, if we took x_2 to be excess of height above 5 feet, say, it would be the weight of a man 5 feet tall. This would, as we all know, change with age. Now the assumption really involved in the scheme of partial regression is that m is constant as respects age, while b is a linear function of age. If this assumption is too simple to agree with obvious facts, the method must be modified. The next simplest assumption would be that m is also a linear function of the age. If we modify the original equation accordingly, it becomes $x_1 = (ax_3 + c)x_2 + dx_3 + e$, that is to say, without going beyond the idea of linearity in the elementary relationships, we find an equation containing a term x_2x_3. Now if we start with an equation containing such a term, and as before make the algebraic sum of the errors zero, the equation takes the form $x_1 = a_{123}[x_2x_3 - r_{23}\sigma_2\sigma_3] + b_{12\cdot3}x_2 + b_{13\cdot2}x_3$, and by the method of least squares, the values of a_{123}, $b_{12\cdot3}$, and $b_{13\cdot2}$ have been determined. For these values the reader is referred to two papers by I. Isserlis in *Biometrika*,[1] which also include an application of a still more complicated equation to data on school children.

Methods of computation. All the r's, m's, and σ's used in the ordinary theory of partial correlation are based directly on the simple r's and σ's whose computation has already been discussed at length. The combination of these simple constants to give the

[1] Vol. x, 1914, pp. 391–411, and vol. xi, 1915, "On the Partial Correlation-Ratio."

Digitized by Google

three-variable constants can be carried out fairly readily on the computing machine, and for that case it is not worth while to learn the special methods which have been elaborated. Even the four-variable m's can be found without any extraction of square roots, according to the formulas given here. The formulas described in the preceding section do require new constants, whose computation will present no great difficulties to those who go into the subject enough to require them.

Summary of Chapter XIV. Complete statistical analysis of a problem often requires a study of the relations between three or more variables.

The first step in such a study is the construction of a series of two-way frequency distributions for the variables taken in pairs.

From these tables the coefficients of correlation, standard deviations, and two-variable regression equations should be found by the methods previously outlined, and preliminary judgments formulated as to which relations are most significant.

By an application of the theory of least squares similar to that used for two variables, the proper coefficients in a three-variable linear equation can be determined. The resulting equation is:

$$X_1/\sigma_1 = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} X_2/\sigma_2 + \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} X_3/\sigma_3$$

The multiple coefficient of correlation, the coefficient of correlation between the observed and the predicted X_1, is given by:

$$R_{1-23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2\,r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

The standard error of the estimates resulting from the regression equation in three variables is given by $\sigma_1\sqrt{(1 - R_{1-23}^2)}$.

The partial coefficient of correlation between two variables measures the correlation between them when the effect of a third variable is considered but eliminated. Its formula is:

$$r_{12-3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)\,(1 - r_{23}^2)}}$$

Digitized by Google

The three-variable regression equation may be represented graphically by a plane.

If we take the sides of a spherical triangle such that their cosines are equal to the ordinary correlation coefficients between the three variables taken in pairs, the cosines of the angles of this spherical triangle will be the partial correlation coefficients and the cosines of the altitudes will be the multiple correlation coefficients.

If good correlation can be found between the items in one historical series and the items as of an earlier date in two other series, the three-variable regression equation provides a valuable instrument for forecasting purposes.

The coefficients in the linear four-variable regression equation may be determined by solving three simultaneous linear equations. The corresponding multiple R may be computed from the ordinary r's (formula long, stated in text).

Regression equations in more than four variables can be treated similarly, but their use is not recommended because of the increasing complexity and the probability that the relationship will fail to be approximately linear.

The simplest non-linear three-variable equation is of the form

$$z = axy + bx + cy + d.$$

According to this equation, the rate of increase of z with x is constant for any particular y, but varies with different values of y, whereas in the linear equation, the rate of increase is the same for all values of y. This equation seems to deserve further study.

Digitized by Google

BIBLIOGRAPHY

If the student who has become familiar with the preceding pages desires to extend his acquaintance with statistical theory, he will find it even more essential to refer to a variety of books and journals than if he had reached a similar stage of progress in almost any other line. One reason for this condition of affairs is the lack of agreement on certain comparatively elementary points, due to the relative youth of Statistics as a separate subject. It is therefore necessary to get the views of two or more writers on such disputed points.

Another reason for consulting various references lies in the fact that statistical methods in different fields have been developed somewhat independently, and economic statisticians, for instance, will endorse crude methods for a problem substantially the same as one in vital statistics for which adequate methods of analysis are available.

A third reason for using a reference list is the idea of some mathematical statisticians that theoretical statistics is merely a part of the mathematical theory of probability, and that therefore the decision as to which of several possible methods is preferable can and should be made on purely mathematical grounds. In the opinion of the reviewer, however, the principle of selection should be, Which method throws most light on a decision which has to be made, either as to some action in the realm of practical affairs, or as to the validity of some suggested law in biometry, psychology, education, economics, or other science? The student will therefore find it desirable, when investigating a new method, to look up both what the mathematical statisticians say about it and how the statisticians in various fields actually use it.

Each of the following references is recommended as worth consulting, but none of them is to be accepted without critical analysis.

JOURNALS.

Journal of the American Statistical Association (prior to January 1, 1922, entitled the *Quarterly Publications of the American Statistical Association*). Published by the American Statistical Association, 474 West 24th Street, New York City. Essential for any American statistician who desires to keep in touch with what contemporaneous statisticians are doing and thinking. Some articles are suggestive rather than final.

Journal of the Royal Statistical Society, published by the Society, London. Contains many noteworthy studies of social, political, and economic

Digitized by Google

subjects, as well as extensions of theory. The discussions of papers presented to the Society are especially informing.

Metron. Tipografia Industrie Grafiche Italiane Padova, Padua, Italy. A journal with articles in English, French, Italian, and other languages, originally announced as intended primarily for the advancement of statistical theory, but actually devoted in large measure to statistical studies in which the method is of only secondary interest.

Biometrika. The University Press, Cambridge, England. Original place of publication of much fundamental statistical theory. Main emphasis originally on biometry, but methods developed have perhaps been used more successfully in educational, economic, and social problems. The reader is expected to have a thorough mathematical training, and to become thoroughly acquainted with the statistical technique developed here and elsewhere by Professor Karl Pearson and his school.

Review of Economic Statistics. Published by the Harvard University Committee on Economic Research, Cambridge, Mass. While primarily concerned with the interpretation of current economic developments, statistical methods for the analysis of historical series are described and discussed, especially in Preliminary Volume I.

SOURCES FOR CURRENT STATISTICAL DATA.

Survey of Current Business. United States Department of Commerce, Washington, D.C. Gives each month the latest figure for each of a large number of important business series, with enough back figures to serve as a basis for comparison.

Publications of Standard Statistics Company, New York. Monthly bulletins give current figures, in some cases a little more promptly than the preceding reference. Annual bulletin gives many series as far back as available.

Statistical Abstract of the United States. Government Printing Office, Washington, D.C. Summarizes important official figures, including Census results, and also some material from other sources.

Statistical Abstract of the United Kingdom. His Majesty's Stationery Office, London, England.

World Almanac and Book of Facts. The New York World, New York. While not official, figures included are generally taken from the best sources. Limitations of space sometimes prevent inclusion of the full explanation of meaning of terms, qualifications, etc., which would be desirable. Consultation of original sources is often wise.

RELATED MATHEMATICAL TOPICS.

Coolidge, J. L. *An Introduction to Mathematical Probability.* Oxford University Press, Oxford, 1925. For students well trained in algebra

Digitized by Google

and calculus, a valuable analysis of the fundamentals of probability. The theory of least squares and the coefficient of correlation are discussed carefully from the mathematical standpoint.

Rietz, H. L., and others. *Handbook of Mathematical Statistics.* Houghton Mifflin Company, Boston, 1924. A collection of chapters on important statistical topics. Chapter III, on "Interpolation, Summation, and Graduation," is a good summary treatment of an important topic omitted from our discussion for lack of space. Chapter XII, on "Index Numbers," may also prove helpful as an alternative presentation of about the same point of view as contained in our Chapter VI. Some of the other chapters are useful.

Running, T. R. *Empirical Formulas.* Wiley, New York, 1917. Gives valuable methods for curve-fitting.

Whittaker, E. T. and Robinson, G. *The Calculus of Observations; a Treatise on Numerical Mathematics.* Blackie & Son, London, 1924. Good chapters on interpolation, normal frequency distributions, method of least squares, practical Fourier analysis, graduation, and correlation, all from the point of view of the worker in the exact sciences who has a thorough mathematical training.

Fisher, Arne. *The Mathematical Theory of Probability.* Second edition. The Macmillan Company, New York, 1922. A valuable treatment of the elements of the theory of probability, some parts of it requiring less mathematical training than the first reference of this group.

GENERAL THEORY OF STATISTICS.

Yule, G. U. *An Introduction to the Theory of Statistics.* C. Griffin & Co., London. First edition, 1910; seventh edition, 1924. Recommended for reference use to accompany this text. A student who has some acquaintance with statistical theory or method will find Yule's discussions, examples, and bibliography very valuable.

Bowley, A. L. *Elements of Statistics.* Fourth edition. P. S. King & Son, London, 1920. Valuable both because of the author's long and distinguished career in economic and governmental investigations and because of some of the discussions of probability in Part II.

Jones, D. C. *A First Course in Statistics.* G. Bell & Sons, London, 1921. Contains a good introductory discussion of the Pearsonian frequency curves.

Kelley, T. L. *Statistical Method.* The Macmillan Company, New York, 1924. Profound and stimulating, but not lucid and sometimes not a safe guide as to practical method.

Zizek, F. *Statistical Averages.* Translated by W. M. Persons. Henry Holt & Co., New York, 1913. Attempts to cover only ratios and averages, but within its range is of great value to the student who wishes to develop sound statistical judgment.

Digitized by Google

Fisher, R. A. *Statistical Methods for Research Workers.* Oliver & Boyd, London, 1925. Explanation of new methods for testing reliability of results derived from small samples.

GRAPHICAL STATISTICAL METHODS.

Brinton, W. C. *Graphic Methods for Presenting Facts.* Engineering Magazine Company, New York, 1914. The pioneer book in this field; is to some extent an elementary text on statistical method from the point of view of the graphic statistician.

Karsten, K. G. *Charts and Graphs.* Prentice-Hall, New York, 1923. A very clear and detailed exposition of graphic methods, including along with it explanation without proofs of many statistical principles. Many of the graphic methods discussed are too complicated for use except by specialists or engineers. The fact that in some points the author is not orthodox increases the attractiveness of the book for some readers, but necessitates caution in using his meanings of terms for a general audience.

Riggleman, J. R. *Graphic Methods for Presenting Business Statistics.* McGraw-Hill, New York, 1926. Teaches the use of graphic devices as a language without discussing, except very incidentally, statistical principles.

ECONOMIC, BUSINESS, AND SOCIAL STATISTICS.

Mills, F. C. *Statistical Methods Applied to Economics and Business.* Henry Holt & Co., New York, 1924. A good text for the business statistician who wants methods rather than complete proofs and full discussion of the methods. For the student of our text, provides a second point of view on many questions, differently focused rather than fundamentally different.

Chaddock, R. E. *Principles and Methods of Statistics.* Houghton Mifflin Company, Boston, 1925. Intended primarily for students of the social sciences with little mathematical training. Brings out with unusual care the meaning of processes used and emphasizes logical points often overlooked even when elaborate technical statistical processes are employed.

VITAL STATISTICS.

Davenport, C. B. *Statistical Methods, with Special Reference to Biological Variation.* Wiley, New York. Second edition, 1904. A handy collection of important formulas and summary of methods, for biometric work of the general type initiated by Karl Pearson and his school. Contains a convenient table of the ordinates and areas of the normal curve, for closely spaced values of the abscissa.

Whipple, G. C. *Vital Statistics.* Wiley, New York, 1922. An able and

Digitized by Google

lucid exposition of statistical methods as applied to the problems which confront the public health officer.

PSYCHOLOGICAL AND EDUCATIONAL STATISTICS.

Brown, W., and Thomson, G. H. *The Essentials of Mental Measurement.* Cambridge University Press, Cambridge, England, 1921. Recognized as standard in the field of psychological measurement. Theoretically important for its discussion of how to deal with magnitudes which can be arranged in ascending or descending order, but cannot be measured in terms of a scale based on ordinary physical or geometric units.

Rugg, H. O. *Statistical Methods Applied to Education.* Houghton Mifflin Company, Boston, 1917. While not complete in the light of recent developments, valuable for its careful discussions and emphasis on the function of statistics as a guide to executive decisions.

Otis, A. S. *Statistical Method in Educational Measurement.* World Book Company, Yonkers, New York, 1925. While "the purpose of this book is to present the subject in such a way that it can be understood by those who know nothing whatever about statistical method," most of the treatment is thoroughly sound and fits in satisfactorily with the more mathematical developments of the present text.

STATISTICAL AND MATHEMATICAL TABLES.

Glover, J. W. *Tables of Applied Mathematics, Finance, Insurance, Statistics.* George Wahr, Ann Arbor, Michigan, 1923. Contains ordinates and areas of various frequency curves, tables to facilitate use of interpolation formulas, as well as logarithms, compound interest functions, mortality tables, and many other valuable tables.

Pearson, Karl. *Tables for Statisticians and Biometricians.* Cambridge University Press, Cambridge, England, 1914. Brings together some of the important additions to the world's stock of tables made or improved by Professor Pearson or his school.

Pearson, Karl. *Tables of the Incomplete Gamma Function.* His Majesty's Stationery Office, London, 1922. Accurate areas and ordinates of the special skew frequency curves called "Gamma" curves in Chapter VII.

Note also the tables included in the *Handbook of Mathematical Statistics,* Davenport, and Kelley.

BIBLIOGRAPHIES.

For additional references, the reader is referred to the *Handbook of Mathematical Statistics,* Yule, and Kelley.

Digitized by Google

APPENDIX I

DECIMAL VALUES OF COMMON FRACTIONS WITH 2-DIGIT DENOMINATORS, TO FIVE SIGNIFICANT FIGURES

Denom.	Numerator				
	10	20	30	40	50
11	.90909	1.8182	2.7273	3.6364	4.5455
12	.83333	1.6667	2.5000	3.3333	4.1667
13	.76923	1.5385	2.3077	3.0769	3.8462
14	.71429	1.4286	2.1429	2.8571	3.5714
15	.66667	1.3333	2.0000	2.6667	3.3333
16	.62500	1.2500	1.8750	2.5000	3.1250
17	.58824	1.1765	1.7647	2.3529	2.9412
18	.55556	1.1111	1.6667	2.2222	2.7778
19	.52632	1.0526	1.5789	2.1053	2.6316
20	.50000	1.0000	1.5000	2.0000	2.5000
21	.47619	.95238	1.4286	1.9048	2.3810
22	.45455	.90909	1.3636	1.8182	2.2727
23	.43478	.86957	1.3043	1.7391	2.1739
24	.41667	.83333	1.2500	1.6667	2.0833
25	.40000	.80000	1.2000	1.6000	2.0000
26	.38462	.76923	1.1538	1.5385	1.9231
27	.37037	.74074	1.1111	1.4815	1.8519
28	.35714	.71429	1.0714	1.4286	1.7857
29	.34483	.68966	1.0345	1.3703	1.7241
30	.33333	.66667	1.0000	1.3333	1.6667
31	.32258	.64516	.96774	1.2903	1.6129
32	.31250	.62500	.93750	1.2500	1.5625
33	.30303	.60606	.90909	1.2121	1.5152
34	.29412	.58824	.88235	1.1765	1.4706
35	.28571	.57143	.85714	1.1429	1.4286
36	.27778	.55556	.83333	1.1111	1.3889
37	.27027	.54054	.81081	1.0811	1.3514
38	.26316	.52632	.78947	1.0526	1.3158
39	.25641	.51282	.76923	1.0256	1.2821
40	.25000	.50000	.75000	1.0000	1.2500
41	.24390	.48780	.73170	.97561	1.2195
42	.23810	.47619	.71429	.95238	1.1905
43	.23256	.46512	.69767	.93023	1.1628
44	.22727	.45455	.68182	.90909	1.1364
45	.22222	.44444	.66667	.88889	1.1111
46	.21739	.43478	.65217	.86957	1.0870
47	.21277	.42553	.63830	.85106	1.0638
48	.20833	.41667	.62500	.83333	1.0417
49	.20408	.40816	.61225	.81633	1.0204
50	.20000	.40000	.60000	.80000	1.0000
51	.19608	.39216	.58824	.78431	.98039
52	.19231	.38462	.57692	.76923	.96154
53	.18868	.37736	.56604	.75472	.94340
54	.18519	.37037	.55556	.74074	.92593

Digitized by Google

DECIMAL VALUES OF COMMON FRACTIONS WITH 2-DIGIT DENOMINATORS, TO FIVE SIGNIFICANT FIGURES (*continued*)

Denom.	Numerator			
	60	70	80	90
11	5.4545	6.3636	7.2727	8.1818
12	5.0000	5.8333	6.6667	7.5000
13	4.6154	5.3846	6.1538	6.9231
14	4.2857	5.0000	5.7143	6.4286
15	4.0000	4.6667	5.3333	6.0000
16	3.7500	4.3750	5.0000	5.6250
17	3.5294	4.1176	4.7059	5.2941
18	3.3333	3.8889	4.4444	5.0000
19	3.1579	3.6842	4.2105	4.7368
20	3.0000	3.5000	4.0000	4.5000
21	2.8571	3.3333	3.8095	4.2857
22	2.7273	3.1818	3.6364	4.0909
23	2.6087	3.0435	3.4783	3.9130
24	2.5000	2.9167	3.3333	3.7500
25	2.4000	2.8000	3.2000	3.6000
26	2.3077	2.6923	3.0769	3.4615
27	2.2222	2.5926	2.9630	3.3333
28	2.1429	2.5000	2.8571	3.2143
29	2.0690	2.4138	2.7586	3.1035
30	2.0000	2.3333	2.6667	3.0000
31	1.9355	2.2581	2.5806	2.9032
32	1.8750	2.1875	2.5000	2.8125
33	1.8182	2.1212	2.4242	2.7273
34	1.7647	2.0588	2.3529	2.6471
35	1.7143	2.0000	2.2857	2.5714
36	1.6667	1.9444	2.2222	2.5000
37	1.6216	1.8919	2.1622	2.4324
38	1.5789	1.8421	2.1053	2.3684
39	1.5385	1.7949	2.0513	2.3077
40	1.5000	1.7500	2.0000	2.2500
41	1.4634	1.7133	1.9512	2.1951
42	1.4286	1.6667	1.9048	2.1429
43	1.3953	1.6279	1.8605	2.0930
44	1.3636	1.5909	1.8182	2.0455
45	1.3333	1.5556	1.7778	2.0000
46	1.3043	1.5217	1.7391	1.9565
47	1.2766	1.4894	1.7021	1.9149
48	1.2500	1.4583	1.6667	1.8750
49	1.2245	1.4286	1.6327	1.8367
50	1.2000	1.4000	1.6000	1.8000
51	1.1765	1.3725	1.5686	1.7647
52	1.1538	1.3462	1.5385	1.7308
53	1.1321	1.3208	1.5094	1.6981
54	1.1111	1.2963	1.4815	1.6667
55	1.0909	2.2727	1.4545	1.6363

Digitized by Google

DECIMAL VALUES OF COMMON FRACTIONS WITH 2-DIGIT DENOMINATORS, TO FIVE SIGNIFICANT FIGURES (*continued*)

Denom.	Numerator				
	10	20	30	40	50
55	.18182	.36364	.54545	.72727	.90909
56	.17857	.35714	.53571	.71429	.89286
57	.17544	.35088	.52632	.70175	.87719
58	.17241	.34483	.51724	.68966	.86207
59	.16949	.33898	.50847	.67797	.84746
60	.16667	.33333	.50000	.66667	.83333
61	.16393	.32787	.49180	.65574	.81967
62	.16129	.32258	.48387	.64516	.80645
63	.15873	.31746	.47619	.63492	.79365
64	.15625	.31250	.46875	.62500	.78125
65	.15385	.30769	.46154	.61538	.76923
66	.15152	.30303	.45455	.60606	.75758
67	.14925	.29851	.44776	.59701	.74627
68	.14706	.29412	.44118	.58824	.73529
69	.14493	.28986	.43478	.57971	.72464
70	.14286	.28571	.42857	.57143.	.71429
71	.14085	.28169	.42254	.56338	.70423
72	.13889	.27778	.41667	.55556	.69444
73	.13699	.27397	.41096	.54795	.68493
74	.13514	.27027	.40541	.54054	.67568
75	.13333	.26667	.40000	.53333	.66667
76	.13158	.26316	.39474	.52632	.65789
77	.12987	.25974	.38961	.51948	.64935
78	.12821	.25641	.38462	.51282	.64103
79	.12658	.25316	.37975	.50633	.63291
80	.12500	.25000	.37500	.50000	.62500
81	.12346	.24691	.37037	.49383	.61728
82	.12195	.24390	.36585	.48780	.60976
83	.12048	.24096	.36145	.48193	.60241
84	.11905	.23810	.35714	.47619	.59524
85	.11765	.23529	.35294	.47059	.58824
86	.11628	.23256	.34884	.46512	.58140
87	.11494	.22989	.34483	.45977	.57471
88	.11364	.22727	.34091	.45455	.56818
89	.11236	.22472	.33708	.44944	.56180
90	.11111	.22222	.33333	.44444	.55556
91	.10989	.21978	.32967	.43956	.54945
92	.10870	.21739	.32609	.43478	.54348
93	.10753	.21505	.32258	.43011	.53763
94	.10638	.21277	.31915	.42553	.53191
95	.10526	.21053	.31579	.42105	.52632
96	.10417	.20833	.31250	.41667	.52083
97	.10309	.20619	.30928	.41237	.51546
98	.10204	.20408	.30612	.40816	.51020
99	.10101	.20202	.30303	.40404	.50505
101	.09901	.19802	.29703	.39604	.49505

Digitized by Google

DECIMAL VALUES OF COMMON FRACTIONS WITH 2-DIGIT DENOMINATORS, TO FIVE SIGNIFICANT FIGURES (*continued*)

Denom.	Numerator			
	60	70	80	90
56	1.0714	1.2500	1.4286	1.6071
57	1.0526	1.2281	1.4035	1.5789
58	1.0345	1.2069	1.3793	1.5517
59	1.0169	1.1864	1.3559	1.5254
60	1.0000	1.1667	1.3333	1.5000
61	.98361	1.1475	1.3115	1.4754
62	.96774	1.1290	1.2903	1.4516
63	.95238	1.1111	1.2698	1.4286
64	.93750	1.0938	1.2500	1.4062
65	.92308	1.0769	1.2308	1.3846
66	.90909	1.0606	1.2121	1.3636
67	.89552	1.0448	1.1940	1.3433
68	.88235	1.0294	1.1765	1.3235
69	.86957	1.0145	1.1594	1.3043
70	.85714	1.0000	1.1429	1.2857
71	.84507	.98592	1.1268	1.2676
72	.83333	.97222	1.1111	1.2500
73	.82192	.95890	1.0959	1.2329
74	.81081	.94595	1.0811	1.2162
75	.80000	.93333	1.0667	1.2000
76	.78947	.92105	1.0526	1.1842
77	.77922	.90909	1.0390	1.1688
78	.76923	.89744	1.0256	1.1538
79	.75949	.88608	1.0127	1.1392
80	.75000	.87500	1.0000	1.1250
81	.74074	.86420	.98765	1.1111
82	.73170	.85366	.97561	1.0976
83	.72289	.84337	.96386	1.0843
84	.71429	.83333	.95238	1.0714
85	.70588	.82353	.94118	1.0588
86	.69767	.81395	.93023	1.0465
87	.68966	.80460	.91954	1.0345
88	.68182	.79545	.90909	1.0227
89	.67416	.78652	.89888	1.0112
90	.66667	.77778	.88889	1.0000
91	.65934	.76923	.87912	.98901
92	.65217	.76087	.86957	.97826
93	.64516	.75269	.86022	.96774
94	.63830	.74468	.85106	.95745
95	.63158	.73684	.84211	.94737
96	.62500	.72917	.83333	.93750
97	.61856	.72165	.82474	.92784
98	.61225	.71429	.81633	.91837
99	.60606	.70707	.80808	.90909
101	.59406	.69307	.79208	.89109

Digitized by Google

How to Use Preceding Table of Decimal Values

Example 1. What per cent is 35 of 53? i.e., find decimal value of 35/53 and multiply by 100.

Solution:

Read from table, in row of figures to right of 53, in column under 30, .56604. This is 30/53.

And in column under 50, note that 50/53 = .94340; so 5/53 = .09434.

Therefore, adding, 35/53 = .66038. 35 is therefore 66.04 per cent of 53.

Note that when we divide .94340 by 10 by shifting the decimal point, we also drop the last digit. As the corresponding digit in .56604 is not stated, we could not be sure of the sixth digit in our sum, even if we retained it in .094340. Note also that when the final result is stated, even the fifth digit is discarded. It could be retained, but may be incorrect by one unit, as the rounding off of each of the terms added may mean an error in each case of nearly 1/2 unit. By rounding off to four places, we are sure we are right, unless the fifth digit is a 5, in which case, if accuracy is important, we must use a computing machine or do as in Example 2.

Example 2. Find 57/67 correct to four significant figures.

$$\begin{aligned} \text{We have} \quad 50/67 &= .74627 \\ 7/67 &= .10448 \\ \text{Total} \quad &= .85075 \end{aligned}$$

It is not clear whether this should be .8508 or .8507 when rounded off. We therefore perform one step in the process of long division, and get

$$\begin{aligned} 57/67 = .8 + (57 - .8 \times 67)/67 &= .8 + 3.4/67 \\ 3/67 &= .044776 \\ .4/67 &= .005970, \text{ by the table.} \\ 3.4/67 &= .050746 \end{aligned}$$

We therefore find .8507 is closer.

Example 3. To find approximately 400/583.

$$\begin{aligned} \text{From table,} \quad 400/580 &= .68966 \\ 400/590 &= .67797 \\ \text{Difference} \quad &.01169 \end{aligned}$$

Now 400/583 is to a first approximation .3 of the way from the first to the second value noted; that is to say, we must subtract .3 of .01169 (= .00351) from .68966, giving as the answer according to this method .68615. The exact answer is .68611, and in general these interpolated results are not to be relied on in the last digit.

The table can thus be used to give results correct within 1 unit in the fifth significant figure, for fractions consisting of two digit numerators and denominators. For three digit numerators, three numbers from the table are added instead of two; otherwise the work is as before. The possible error in this case is 3/2 unit in the fifth significant figure.

For three digit denominators, a close approximation can be found by inter-

Digitized by Google

polation. In general, for three or more digit denominators, an estimate of the result can be made by inspection from this table; if a rigorously exact value is wanted, use a computing machine or long division, as the interpolation is in such cases nearly as laborious as the latter, and less accurate. Another scheme is to use a table of reciprocals, and then one of the larger multiplication tables. This method will not appeal to many, however.

Practice Examples in the Use of the Table

Example	Answer
Reduce 43/97 to a decimal fraction	.44330 ± .00001
What per cent of 49 is 17?	35.694 ± .001
Reduce 527/38 to a decimal	13.868 ± .0015
Find the whole number in the preceding by division, and the fraction by the table	13.86842 ± .00001
Find the best four digit value of 65/38 from the table	1.711
What per cent is 43 of 538?	7.9930 by table 7.9926 by division

Show algebraically that in all cases in which linear interpolation is used to secure approximate values of fractions with 3-digit denominators, the true value will be less than this approximate value.

Digitized by Google

APPENDIX II

DENOMINATORS OF a AND b', COEFFICIENTS OF PARABOLIC REGRESSION FORMULA AS APPLIED TO HISTORICAL SERIES

N	Denominator of b' i.e., $\frac{N(N^2-1)}{12}$	Denominator of a i.e., $\frac{N(N^2-1)(N^2-4)}{180}$
2	0.5	
3	2	.67
4	5	4
5	10	14
6	17.5	37.33
7	28	84
8	42	168
9	60	308
10	82.5	528
11	110	858
12	143	1,334.67
13	182	1,992
14	227.5	2,912
15	280	4,125.33
16	340	5,712
17	408	7,752
18	484.5	10,336
19	570	13,566
20	665	17,556
21	770	22,432.67
22	885.5	28,336
23	1,012	35,420
24	1,150	43,853.33
25	1,300	53,820
26	1,462.5	65,520
27	1,638	79,170
28	1,827	95,004
29	2,030	113,274
30	2,247.5	134,250.67
31	2,480	158,224
32	2,728	185,504
33	2,992	216,421.33
34	3,272.5	251,328
35	3,570	290,598
36	3,885	334,628
37	4,218	383,838
38	4,569.5	438,672
39	4,940	499,598.67
40	5,330	567,112

Digitized by Google

DENOMINATORS OF a AND b', COEFFICIENTS OF PARABOLIC REGRESSION FORMULA AS APPLIED TO HISTORICAL SERIES (*continued*)

N	Denominator of b' i.e., $\frac{N(N^2-1)}{12}$	Denominator of a i.e., $\frac{N(N^2-1)(N^2-4)}{180}$
41	5,740	641,732
42	6,170.5	724,005.33
43	6,622	814,506
44	7,095	913,836
45	7,590	1,022,626
46	8,107.5	1,141,536
47	8,648	1,271,256
48	9,212	1,412,506.67
49	9,800	1,565,040
50	10,412.5	1,732,640
51	11,050	1,913,123.33
52	11,713	2,108,340
53	12,402	2,319,174
54	13,117.5	2,546,544
55	13,860	2,791,404
56	14,630	3,054,744
57	15,428	3,337,590.67
58	16,254.5	3,641,008
59	17,110	3,966,098
60	17,995	4,314,001.33

Digitized by Google

APPENDIX III

ORDINATES AND AREAS OF CERTAIN THEORETICAL FREQUENCY CURVES

Equation of gamma curves $y = \frac{e^{-n}n^{n-1}}{\Gamma(n)}e^{-\sqrt{n}x}\left(1+\frac{x}{\sqrt{n}}\right)^{n-1}$

Equation of normal curve $y = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$

The areas and ordinates of the normal curve are quoted from *Tables for Statisticians and Biometricians*, edited by Professor Karl Pearson and published by the Cambridge University Press. The reader is referred to those tables for additional values.

The areas of the gamma curves are derived, with some slight modification in origin and scale, from the *Tables of the Incomplete Gamma Function*, also prepared by Professor Karl Pearson and his staff.

x	Gamma Curve, $n = 4$		Gamma Curve, $n = 9$		Gamma Curve, $n = 16$		Normal Curve	
	Ordinate	Area	Ordinate	Area	Ordinate	Area	Ordinate	Area
−4					.0000	0.0	.0001	.00003
−3				.00000	.0001	.00000	.0044	.00135
−2.5			.0004	.00003	.0036	.00051	.0175	.00621
−2.2			.0074	.00086	.0165	.00315	.0355	.01390
−2	.0000	.00000	.0243	.00380	.0361	.00823	.0540	.02275
−1.9	.0022	.00006	.0386	.00691	.0503	.01253	.0656	.02872
−1.8	.0143	.00078	.0574	.01167	.0678	.01840	.0790	.03593
−1.7	.0395	.00336	.0806	.01853	.0885	.02619	.0940	.04457
−1.6	.0767	.00908	.1080	.02793	.1123	.03620	.1109	.05480
−1.5	.1226	.01899	.1390	.04026	.1389	.04874	.1295	.06671
−1.4	.1735	.03377	.1726	.05582	.1677	.06405	.1497	.08076
−1.3	.2256	.05373	.2076	.07482	.1979	.08232	.1714	.09680
−1.2	.2757	.07881	.2430	.09735	.2289	.10366	.1942	.11507
−1.1	.3213	.10871	.2774	.12338	.2598	.12810	.2179	.13567
−1	.3609	.14288	.3100	.15276	.2896	.15558	.2420	.15866

Digitized by Google

ORDINATES AND AREAS OF CERTAIN THEORETICAL FREQUENCY CURVES (*continued*)

x	Gamma Curve, $n = 4$		Gamma Curve, $n = 9$		Gamma Curve, $n = 16$		Normal Curve	
	Ordinate	Area	Ordinate	Area	Ordinate	Area	Ordinate	Area
−.9	.3933	.18065	.3391	.18523	.3174	.18595	.2661	.18406
−.8	.4180	.22128	.3644	.22044	.3426	.21898	.2897	.21186
−.7	.4351	.26400	.3853	.25797	.3643	.25435	.3123	.24196
−.6	.4450	.30806	.4012	.29733	.3822	.29171	.3332	.27425
−.5	.4481	.35277	.4120	.33803	.3957	.33064	.3521	.30854
−.4	.4452	.39748	.4177	.37956	.4047	.37070	.3683	.34458
−.3	.4372	.44164	.4185	.42141	.4092	.41143	.3814	.38209
−.2	.4249	.48478	.4147	.46311	.4092	.45239	.3910	.42074
−.1	.4092	.52652	.4068	.50422	.4050	.49313	.3970	.46017
0	.3907	.56653	.3953	.54435	.3969	.53326	.3989	.50000
.1	.3704	.60460	.3807	.58317	.3853	.57239	.3970	.53983
.2	.3486	.64055	.3635	.62039	.3706	.61021	.3910	.57926
.3	.3261	.67429	.3445	.65581	.3537	.64645	.3814	.61791
.4	.3034	.70577	.3240	.68924	.3347	.68089	.3683	.65542
.5	.2808	.73497	.3027	.72059	.3143	.71335	.3521	.69146
.6	.2586	.76193	.2809	.74977	.2930	.74372	.3332	.72575
.7	.2371	.78671	.2591	.77677	.2711	.77192	.3123	.75804
.8	.2165	.80938	.2376	.80161	.2492	.79794	.2897	.78814
.9	.1969	.83004	.2167	.82432	.2276	.82178	.2661	.81594
1.0	.1785	.84880	.1966	.84497	.2066	.84349	.2420	.84134
1.1	.1612	.86577	.1774	.86366	.1864	.86313	.2179	.86433
1.2	.1452	.88108	.1594	.88049	.1672	.88080	.1942	.88493
1.3	.1304	.89485	.1425	.89558	.1491	.89660	.1714	.90320
1.4	.1167	.90719	.1269	.90904	.1323	.91066	.1497	.91924
1.5	.1043	.91823	.1125	.92100	.1168	.92311	.1295	.93319

Digitized by Google

ORDINATES AND AREAS OF CERTAIN THEORETICAL FREQUENCY CURVES (*continued*)

x	Gamma Curve, $n = 4$		Gamma Curve, $n = 9$		Gamma Curve, $n = 16$		Normal Curve	
	Ordinate	Area	Ordinate	Area	Ordinate	Area	Ordinate	Area
1.6	.0929	.92808	.0994	.93159	.1026	.93407	.1109	.94520
1.7	.0825	.93685	.0875	.94092	.0897	.94367	.0940	.95543
1.8	.0732	.94463	.0767	.94912	.0780	.95204	.0790	.96407
1.9	.0647	.95152	.0670	.95630	.0676	.95931	.0656	.97128
2.0	.0573	.95762	.0583	.96255	.0583	.96560	.0540	.97725
2.2	.0444	.96774	.0438	.97271	.0428	.97565	.0355	.98610
2.5	.0300	.97877	.0279	.98331	.0262	.98583	.0175	.99379
3.0	.0151	.98966	.0125	.99294	.0108	.99457	.0044	.99865
4.0	.0035	.99771	.0021	.99889	.0015	.99934	.0001	.99997
5.0	.0008	.99953	.0003	.99985	.0002	.99993		.9999997

Digitized by Google

INDEX

A.D., symbol for average deviation, 137.
Adjustment of differences, method of, in regression or trend lines, 157.
Algebraic discussion of chain index numbers, 127; of composite percentages, 45; of effect of differences in weights, 46.
Algebraic method, in regression or trend lines, 157.
Algebraic proof of correctness, of method of combining percentages, 43.
Analysis of time series, 177–96; three different aspects of problem of seasonal variation, 178; summary of methods used to determine seasonal factors, 179; the method of monthly means, 180; the method of monthly per cent of annual, 182; the geometric-mean link-relative method, 185; the median or arithmetic-mean link-relative methods, 188; the method of typical ratio to twelve months' moving average, 189; method of typical ratio to trend expressed by an equation, 190; method of typical ratio to a free-hand curve, 191; method of monthly per cent of annual, modified to eliminate trend, 191; which method to use, 193; changing seasonal factors, 193; seasonals not extremely accurate, 194; summary, 194.
Analysis, percentage, 16 f.
Applications of formulas, in regression or trend lines, 166; of index numbers, 132.
Approach, statistical, 1 f; essential points in, 14.
Arithmetic mean, definition of, 81; relative size, 101; computation by deviation method, 83; effect of a change in weights, 84.
Arithmetic-mean link-relative method, in determining seasonal factors, 188.
Autocatalytic curve, the, 174.
Average, the, 5.
Average deviation, the, 137.
Average failure to correspond, 207.
Average index number, 111.
Average product-sum, 209.
Averages, comparison of, 100; computation of, 104; effect of extreme items, 102; the method of, 203; special, 98.
Averages and their use, 80–107; what an average does not do, 80; definition of arithmetic mean, and elementary formulas, 81; the median, 85; the mode, 88; relation of mode to mean and median, 89; illustration of mean, median and mode from the mortality table, 89; the harmonic mean, 90; illustrations of use of harmonic mean, 92; general principles of averaging ratios, 94; geometric mean, 96; special averages, 98; comparison of the averages, 100; relative size of A.M., G.M., and H.M., 101; effect of extreme items on the different averages, 102; ease of computation and manipulation, 104; which average should be used? 105; summary, 105.
Averaging of ratios, the, 94.
"Averaging in," a percentage based on additional cases, 44.

Background statistics, 13.
Bar graphs, 21.
Binomial distribution, position of mode of, 240; proof and application of formulas, 244, 245, 248.
Binomial expansion and the normal curve, 238–56; algebraic proof of relation to probabilities, 238; position of mode of binomial distribution, 240; effect of increasing n, 243, 249; proof of formulas, 244; applications of formulas, 245, 248; verification of the theory of sampling, 247; justification of the theory of least squares, 252; summary, 254.
Birth rate, 27.
Boundaries of classes, exact definition necessary, 12.

Cells, nine, table with, 202.
Chain index numbers, algebraic discussion of, 127.
Chain method of construction, of index numbers, 125.
Change in weights, effect on average of, 84; effect in combining ratios, 45.

Digitized by Google

Characteristic, typical, 8.
Characteristics, methods of finding relation between, 197–202.
Chart, ratio, 33 f. *See also* Diagrams.
Classes, exact definition of boundaries necessary, 12.
Classification, the principle of, 3; concealed, 49.
Coefficient, rank-difference, 223.
Coefficient of correlation, the, 204–16; methods of derivation of formula, 204–07; probability basis for, 263; multiple, 273; partial, 274; in three-variable relationship, 268.
Coefficient of regression, 209–16.
Coefficient of variation, the, 150.
Combination of ratios, 42–52; effect of changing weights in, 45.
Comparison of averages, 100.
Complication, first, in index numbers, 113, 118; second, 118; third, 123; fourth, 124.
Composite percentages, algebraic discussion of, 45.
Computing machines, 19.
Concealed classification, 49.
Constants, determination in quadratic formula, 162.
Constants, statistical, formulas for standard errors of, 262.
Contour lines, method of, 203.
Coöperation, statistical, 14.
Correlation, the index of, 222; two-variable, 197–226.
Correlation ratio, the, 219.
Cumulative diagrams, 61.
Cumulative graphs with warped scale base, 65.
Current ratio, 28.

Death rate, 27.
Deciles, the, 151.
Denominators, of constants in quadratic regression equation, formulas for, 164.
Deviation, average, 137; mean, 137; general formula, in frequency distribution, 141; a minimum about the median, 143; quartile, 136; standard, 144.
Deviation method, computation by, 83.
Deviations, standardized, 204.
Diagrams, bar, 22; "pie," 23; semi-logarithmic, 33; histogram, 55; frequency polygon, 56; cumulative, 61; "more than," 63; "less than," 63; with warped scale base, 65 f.
Difference of two means, probable error of, 261.
Differences, adjustment of, method in regression or trend lines, 157.
Differences in weights, algebraic discussion of effect, 46.
Dispersion, measures of, 135–54; the range, 135; the quartile deviation, 136; mean deviation, 137; general formula for mean deviation of a frequency distribution, 141; the mean deviation is a minimum about the median, 143; the standard deviation, 144; relations between Q, $A.D.$, and σ, 147; use of σ as a standard, 148; proportion of cases within various distances of mean, 149; other measures of dispersion, 150; instances of use of measures of dispersion, 151; summary, 152.
Distribution, binomial, position of mode of, 240.
Distribution, frequency, 9; general formula for mean deviation of, 141; frequency, two-way, 197–226; quantity, 9; graphical representation of, 69.
Distributions, frequency and quantity, 53–78.
Division tables, the use of, 20.

Equation, of regression or trend lines, general remarks on selection, 174.
Equation of trend, 159.
Error of grouping, 215.
Errors, common, in the use of percentages, 39.
Extreme items, effect on different averages, 102.

Factors, seasonal, changing, 193; methods used to determine, 179–93.
Forecasting the future, use of trend lines in, 169.
Formation of series, importance of process, 11.
Four-variable correlation, 278.
Free-hand curve, in analysis of time series, 191.
Frequencies, percentage, 82.
Frequency distribution, 9, 53–69, 72–79; the histogram, 55; the frequency polygon, 56; cumulative diagrams, 61; "more than" diagrams, 63; cumulative graphs with warped scale base, 65; which methods should be used? 72; summary, 72; general formula for mean

Digitized by Google

deviation of, 141; two-way, 197–226.
Frequency polygon, the, 56.
Frequency table, two-way, 199.
Functional quantitative series, 11; trend lines for, 171.
Future, use of trend lines in forecasting, 169.

Gauss-La Place law of frequency of error, 251.
Geographic series, 6.
Geometric interpretation of multiple and partial correlation, 275.
Geometric mean, 96; relative size, 101; use of, 130.
Geometric-mean link-relative method, in determining seasonal factors, 185.
Graphic method, in regression or trend lines, 156.
Graphic treatment of partition percentages, 21.
Graphical devices for frequency distributions, 73.
Graphical representation of quantity distributions, 69.
Graphs, cumulative, with warped scale base, 65.

Harmonic mean, 90–96; relative size, 101.
Histogram, the, 55.
Historical series, 8, 35; trend lines for, 155–71; analysis of, 177–96; application of multiple correlation to, 277; comparing the lag of, 216.
"Homogeneous" material, 49.

Ideal index number, 117.
Inadequately explained rate, instance of, 28.
Index of correlation, the, 222.
Index numbers, 32, 108–34; first method of solution of typical problem, 108; second method, 109; definition of index number, 110; noteworthy points, 111; first complication, 113; discussion of various methods, 116; importance of first complication, 118; second complication, 118; use of price index numbers to eliminate effect of price change from sales, 120; third complication, 123; fourth complication, 124; chain method of construction, 125; algebraic discussion of chain index numbers, 127; use of the geometric mean, 130; applications of, 132; summary, 132.
Indices, seasonal, the construction of, 186.
Intelligence quotient, 30.
Interest rate, 27.
Interpolation, 38.
Interquartile range, the, 136.

"J-shaped" polygons, 59.

Labor turnover, 28.
Lag, of historical series, comparing one with another, 216.
Least squares, justification of the theory, 252.
"Less than" diagrams, 63.
Linearity, of regression equations, measures of, 220.
Lines, regression or trend, 155–76.
Link-relative methods, in analysis of time series, 185, 188.
Logarithm tables, the use of, 20.
Long hand division, computation by, 20.
Lower quartile, the, 136.

Manipulation of averages, ease of, 104.
Mean, derivation of formula for probable error of, 258; relation to mode and median, 89; relative size of arithmetic, geometric, and harmonic, 101; arithmetic, 81; geometric, 96; use of, 130; harmonic, 90; unweighted, 82.
Mean deviation, 137; of frequency distribution, general formula for, 141.
Means, monthly, the method of, 180; probable error of the difference of two, 261.
Measures of dispersion, 135–54.
Median, the, 85; relation to mean and mode, 89.
Median or arithmetic-mean link-relative methods, in determining seasonal factors, 188.
Mode, the, 88; relation to mean and median, 89; of binomial distribution, position of, 240.
Monthly means, method in analysis of time series, 180; the trend of, 180.
Monthly per cent of annual, 182; method modified to eliminate trend, 191.
"More than" diagrams, 63.
Multiple coefficient of correlation, 273.
Multiple correlation, geometric interpretation of, 275; application to historical series, 277.

Digitized by Google

Multiplication tables the, use of, 20.
n, effect of increasing, in binomial expansion, 243, 249.
Non-linear three-variable regression equation, 279.
Normal curve, the, 60; and binomial expansion, 238–56.

Ogives, 63.
Operating ratio, 28.

p, symbol for average product-sum, 209.
Parallel bars, 23, 24.
Partial correlation, geometric interpretation of, 275; coefficients, 274.
Partition percentages, 16; graphic treatment of, 2.
Percentage, "averaging in," 44; use to measure progress relative to a standard, 24.
Percentage analysis, 16–41.
Percentage frequencies, 82.
Percentage method of analysis, in three-variable relationships, 269.
Percentage weights, 82.
Percentages, advantages of use, 18; computation of, 18; common errors in use of, 39; correct method for combining, 42.
Percentages, composite, algebraic discussion of, 45.
Percentages, partition, 16; graphic treatment of, 21;
Pie diagram, the, 22, 23.
Points, quartile, 136.
Polygon, frequency, 56.
Price index numbers, 108–113, 118–134, use to eliminate effect of price change from sales, 120.
Price level index number, 119.
Probability, sampling and, 229; algebraic proof of, 238.
Probability basis, for the coefficient of correlation, 263.
Probable error, 148, 257–64; meaning of, 257; derivation of formula for probable error of mean, 258; probable error of the difference of two means, 261; formulas for standard errors of statistical constants, 262; a probability basis for the coefficient of correlation, 263; summary, 264.
Progress, use of percentage in measuring, 24.
Protogamic surface, the, 203.

Q, meaning of, 136.
Quadratic formula, determination of constants in, 162.
Qualitative statistical series, 4.
Quantitative series, functional, trend lines for, 171.
Quantitative statistical series, 4.
Quantities, variation in, 113, 118.
Quantity distribution, 9, 53–78; graphical representation of, 69.
Quartile deviation, the, 136.
Quintiles, the, 151.

r, symbol for coefficient of correlation, 205 f.
Range, the, 135.
Rank-difference coefficient of correlation, the, 223; computation of, 225.
Rank list, the, 6.
Rank lists, comparison of, 223.
Ratio, correlation, 219.
Ratio, typical, methods in determining seasonal factors, 189, 190, 191.
Ratio or percentage analysis, 16–41; partition percentages, 16; advantages of the use of percentages, 18; computation of percentages, 18; graphic treatment of partition percentages, 21; use of percentage to measure progress relative to a standard, 24; interest rate, death rate and birth rate, 27; instance of an inadequately explained rate, 28; ratios which are not percentages, 29; relative numbers, 31; the ratio chart, or semi-logarithmic diagram, 33; principle of the ratio chart, 36; common errors in the use of percentages, 39; summary, 39.
Ratio chart, 33, 34; the principle of, 36.
Ratios which are not percentages, 29; the averaging of, 94; between quartiles, average deviation and standard deviation, 147; the combination of, 42–52; a correct method of combining percentages, 42; algebraic proof of this method, 43; "averaging in" a percentage based on additional cases, 44; effect of changing weights in combining ratios, 45; algebraic discussion of composite percentages, 45; algebraic discussion of effect of differences in weights, 46; "homogeneous" material, 49; summary, 51.
Regression coefficient, 159; computation of, 209.

Digitized by Google

Regression equation, 159, 204–16; measures of linearity, 220; non-linear three-variable, 279; in three-variable relationships, 268, 269, 275.
Regression or trend lines, 155–76; method of selected values, 155; graphic method, 156; method of adjustment of differences, 157; algebraic method, 157; methods to be used if straight line is a poor fit, 161; determination of constants in quadratic formula, 162; formulas for denominators, 164; application of formulas, 166; use of trend lines in forecasting the future, 169; trend lines for functional quantitative series, 171; summary, 174.
Relative numbers, 31, 111.
"Root-mean-square," the, 144.

Sales, volume of, procedure in obtaining, 121.
Sampling, elementary ideas in regard to, 229–37; sampling and probability, 229; relation of sampling to ordinary thought and to scientific induction, 230; fundamental points in the theory, 232; summary, 236; verification of theory, 247.
Scatter diagram, the, 198.
Seasonal factors, changing, 193; summary of methods used to determine, 179–93.
Seasonal variation, three different aspects of problem, 178.
Seasonals, degree of accuracy in, 194.
Selected values, in regression or trend lines, 155.
Selection of equation, general remarks on selection for regression or trend lines, 174.
Semi-interquartile range, the, 136.
Semi-logarithmic diagram, the, 33.
Series, importance of process of formation, 11; functional quantitative, 171; geographic, 6; historical, 8; statistical, the genesis of, 2; qualitative, 4; quantitative, 4; essential points in construction, 4; types of, 5; geographic, 6.
Sigma, the symbol, 144.
Σ, meaning of, 46.
Significant digit, 215.
Size-ratios, 29, 80.
Skewness, 60.
Slide rules, 20.
Special averages, 98.
Standard deviation, the, 144; computation of, 211; in three variable relationship, 268; use as a standard, 148.
Standard errors of statistical constants, formulas for, 262.
Standardized deviations, 204.
Statistical constants, formulas for standard errors of, 262.
Statistical series, the genesis of, 2; types of, 5.
Statistics, meaning of, 1–15; methods, 1; the genesis of statistical series, 2; types of statistical series, 5; importance of the process of formation of series, 11; exact definition of boundaries of classes necessary, 12; background statistics, 13; statistical cooperation, 14; essential points in the statistical approach, 14.
Straight line, equation of, 157.

Tabulation, the device of, 3.
Tax rate, 28.
Tertiles, the, 151.
Three- and four-variable correlation, 265–82; a problem in three-variable relationship, 265; coefficients of correlation, 268; standard deviations, 268; regression equations, 268; discussion of equations and correlations, 268; application of percentage method of analysis, 269; regression equation in three variables, 269; multiple coefficient of correlation, 273; partial correlation coefficients, 274; three-variable regression equation, 275; geometric interpretation of multiple and partial correlation, 275; application of multiple correlation to historical series, 277; four-variable correlation, 278; non-linear three-variable regression equation, 279; methods of computation, 280; summary, 281.
Time series, analysis of, 177–96.
Trend lines, 155–76.
Two-way frequency distributions, 197–226; the method of averages, 203; the coefficient of correlation, 204; magnitude of r, 205; second method of derivation of formula for coefficient of correlation, 206; significance of various values of r, 207; computation of coefficients of correlation and regression, 209; computation of r for a simple case of pairs of values of x and y, 210; computation of coefficients of regression

Digitized by Google

and correlation from two-way frequency table, 210; use of r for determining lag of one historical series as compared with another, 216; the correlation ratio, 219; measures of linearity of the regression equations, 220; the index of correlation, 222; comparison of rank lists, and the rank-difference coefficient, 223; example showing computation of rank-difference coefficient of correlation, 225; summary, 227.
Two-way frequency table, computation of coefficients of regression and correlation from, 210.
Typical characteristic, 8.

Unweighted mean, 82.
Upper quartile, the, 136.

Values, selected, method in regression or trend lines, 155.
Variation, coefficient of, 150.
Variation, seasonal, three different aspects of problem, 178.
Variation in quantities, in constructing index numbers, 113, 118.

Wages, hourly rates, diagram of, 9.
Warped scale base, cumulative graphs with, 65.
Weights, 43, 84; effect of changing in combining ratios, 45; algebraic discussion of effect of differences, 46; percentage, 82.

Digitized by Google

Digitized by Google

Digitized by Google

Digitized by Google

14 DAY USE
RETURN TO DESK FROM WHICH BORROWED

LOAN DEPT.

This book is due on the last date stamped below, or on the date to which renewed.
Renewed books are subject to immediate recall.

2 NOV'64 MS	
REC'D LD	
OCT 19'64-10 AM	
8Jan'65LMX	
AUG 14 1969 06	
REC'D LD	AUG 13'69-2 PM

LD 21A–40*m*-11,'63
(E1602s10)476B

General Library
University of California
Berkeley

Digitized by Google

YC 05217

Digitized by Google

14 DAY USE
RETURN TO DESK FROM WHICH BORROWED

LOAN DEPT.

This book is due on the last date stamped below, or
on the date to which renewed.
Renewed books are subject to immediate recall.

2 NOV'64 MS

REC'D LD

OCT 19'64-10 AM

8Jan'65LMX

REC'D LD

JAN [illegible] 1 PM

AUG 14 1969 06

REC'D LD AUG 13'69-2 PM

LD 21A-40m-11,'63
(E1602s10)476B

General Library
University of California
Berkeley

Digitized by Google

YC 05217

Digitized by Google